Reconocimiento de acciones cotidianas

Kelly Vizconde La Motta

Orientador: Dr.Guillermo Cámara Chávez

Jurado:
Dr. David Menotti – Universidade Federal do Paraná – Brasil
Dr. Juan Carlos Gutierrez – Universidad Católica San Pablo – Perú
Dr. Erick Gomez Nieto – Universidade de Sao Paulo – Brasil
Dr. Alex Cuadros – Universidad Católica San Pablo – Perú

Tesis presentada al
Centro de Investigación e Innovación en Ciencia de la Computación (RICS)
como parte de los requisitos para obtener el grado de
Maestro en Ciencia de la Computación.

Universidad Católica San Pablo – UCSP
Noviembre de 2016 – Arequipa – Perú

A Dios, por todo lo que me ha dado, a todos los profesores por sus enseñanzas y especialmente a mi familia por su apoyo incondicional.

Agradecimientos

En primer lugar deseo agradecer a Dios por haberme guiado a lo largo de estos x años de estudio.

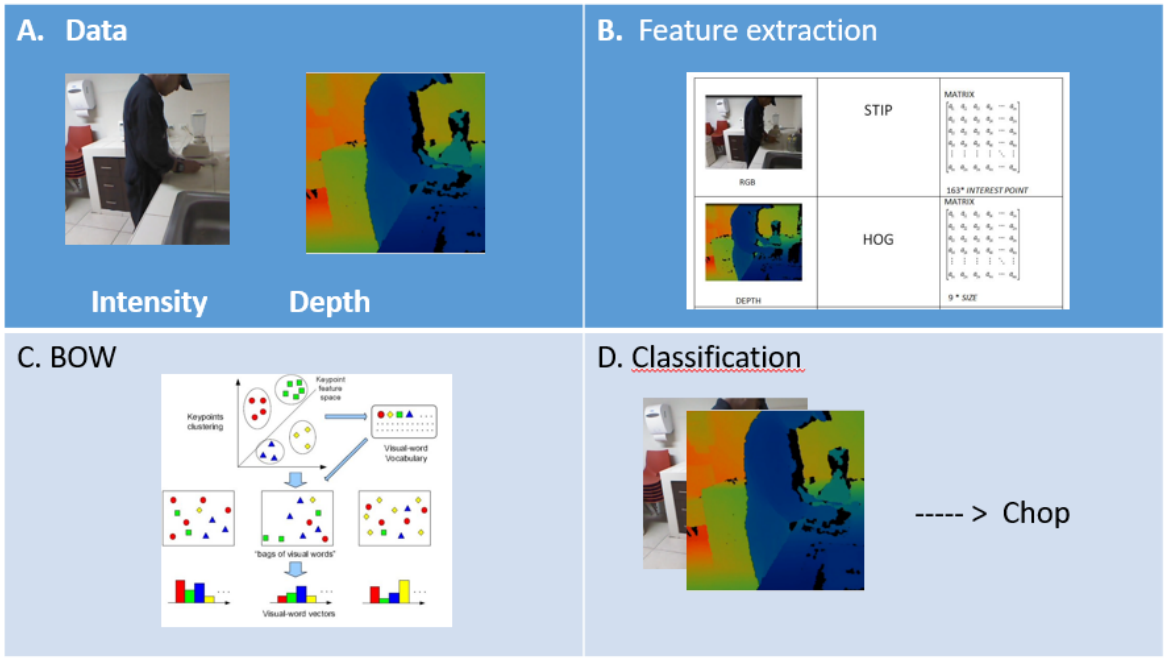
Agradezco a mi familia, a mi papá quien siempre me brindo su apoyo incondicional a pesar de que no siempre estuvimos de acuerdo, mi mamá quien siempre llevaba una manta al momento de hacer tareas de amanecida, porque el trabajo no me permitía hacerlo en otra hora, mi hermana quien a veces trabajó el doble de horas para que pueda asistir a clases y nunca me dejó flaquear, a Helena e Icker quienes hacían tanta bulla que no dejaban dormir a pesar del cansancio.

Agradezco a la universidad, al Dr. Alex Cuadros quien confió en mí y me brindó la oportunidad de estudiar la maestría y especialmente a mi orientador el Dr. Guillermo Camara por haberme guiado en esta tesis, haber tenido paciencia conmigo y agradecerle el haberme convertido en una profesional de calidad.

Deseo agradecer de manera especial al Consejo Nacional de Ciencia, Tecnología e Innovación Tecnológica (CONCYTEC) y al Fondo Nacional de Desarrollo Científico, Tecnológico e Innovación Tecnológica (FONDECYT-CIENCIACTIVA), que mediante Convenio de Gestión UCSP-FONDECYT 011-2013, han permitido la subvención y financiamiento de mis estudios de Maestría en Ciencia de la Computación en la Universidad Católica San Pablo (UCSP).

Deseo agradecer también a mis compañeros, de todos y cada uno de ellos aprendí algo nuevo y desearles lo mejor en su vida profesional.

Abstract



The proposed method consists of three parts: features extraction, the use of bag of words and classification. For the first stage, we use the STIP descriptor for the intensity channel and HOG descriptor for the depth channel, MFCC and Spectrogram for the audio channel. In the next stage, it was used the bag of words approach in each type of information separately. We use the K-means algorithm to generate the dictionary. Finally, a SVM classifier labels the visual word histograms. For the experiments, we manually segmented the videos in clips containing a single action, achieving a recognition rate of 94.4 % on Kitchen-UCSP dataset, our own dataset and a recognition rate of 88 % on HMA videos.

Keywords: STIP, HOG, Spectogram, SVM, Bag of Words.

Resumen

El método propuesto consta de tres partes: la extracción de características, el uso de bolsa de palabras y la clasificación. Para la primera etapa se usó los descriptores STIP para el canal de intensidad, HOG para el canal de profundidad , MFCC y Espectrograma para el canal de audio. En la siguiente etapa se utilizó bolsa de palabras en cada tipo de información por separado. Para la generación del diccionario se usó K-means y para el proceso de clasificación se utilizó SVM. En la parte de experimentos los vídeos fueron divididos en clips, llegando a tener una tasa de asertividad del 94.4 % en la base de vídeos Kitchen-UCSP, que fue elaborada para esta investigación y una tasa de asertividad del 88 % en la base de vídeos HMA.

Palabras clave:STIP, HOG, Espectograma, SVM, Bolsa de palabras.

Índice general

Índice de cuadros

Índice de figuras

Capítulo 1

Introducción

El reconocimiento de acciones humanas es un tema importante dentro del área de visión por computador; lograr que una máquina pueda interpretar y reconocer por sí sola una acción, sin la intervención de un humano, es el motivo de esta y de varias investigaciones. Sus aplicaciones involucran interacciones entre personas y dispositivos, tales como interfaces hombre-máquina. La mayoría de estas aplicaciones requieren un reconocimiento automático de las actividades de alto nivel, logrando varios beneficios en el ámbito donde se apliquen, por ejemplo:

- **Sistemas interactivos**: Robots sociales que comparten un espacio con la gente, requieren la capacidad para detectar y rastrear humanos. Este conocimiento es clave para la integración efectiva de los robots en un ambiente humano; como lo hace (Stork et al., 2012) donde el reconocimiento de actividades mediante la información auditiva logra la descripción de acciones a personas.

- **Video-Vigilancia**: El poder reconocer una acción humana dentro de bancos, aeropuertos, fronteras, o en espacios públicos ayuda con la seguridad de una ciudad o donde se aplique.

- **Salud**: Sistemas de protección de personas mayores y/o niños, fisioterapia asistida por ordenadores, ordenadores que hacen la labor de entrenadores en gimnasios, análisis semántico de movimientos son algunas de las áreas donde el reconocimiento de acciones ha dado grandes avances.

El ser humano puede analizar e interpretar datos visuales de forma rápida y sencilla, asi mismo superar sin ninguna dificultad adversidades impuestas por el ambiente (*e.g.* variaciones de tamaño, posición, oclusión, variación en las condiciones de iluminación, diferentes posturas del cuerpo, diferentes ángulos de visualización, variabibilidad en el fondo de la imagen).

1.1. Justificación

La mayoría de trabajos basan sus propuestas en datos visuales (RGB) para el reconocimiento de acciones, es importante resaltar que el análisis de vídeos es intrínsecamente multimodal, exigiendo un conocimiento multidisciplinario. Trabajos anteriores en reconocimiento de acciones han dado énfasis al uso de descriptores locales (Laptev, 2005; Alcântara et al., 2014; Wang et al., 2009) demostraron que no existe un descriptor de características que sea optimo para todas las bases de datos. El canal de intensidad (RGB) es vulnerable a las variaciones de iluminación y fondo, por lo que la pérdida de información es significativa, reduciendo asi la capacidad de los descriptores. La aparición del sensor Kinect^{TM} revolucionó el campo de visión por computador, brindando mapas de profundidad a bajo costo; como los sensores de profundidad son relativamente nuevos, la extracción de características a partir de este tipo de datos adaptan ligeramente las mismas técnicas de extracción usadas en el dominio RGB. Uno de los problemas que pudo superar el canal de profundidad a comparación del canal de intensidad es la vulnerabilidad a la variación de luz, otra ventaja que brinda es la fácil segmentación que esta produce, por lo que las oclusiones parciales pueden ser superadas.

Sin embargo, estas modalidades se limitan al campo de visión de la imagen por lo que no es robusto en todos los rangos de condiciones ambientales como se observa en la Figura 1.1. Por otra parte, la información visual no siempre puede proporcionar evidencia acerca de las acciones, por lo que se opta tomar una percepción auditiva, ya que muchas actividades humanas producen sonidos muy característicos, lo que infiere de manera efectiva las acciones humanas correspondientes.

Figura 1.1: Problemas de iluminación, tamaño, posición, fondo de imagen y oclusión son unas de las limitaciones y condiciones ambientales que se presentan.

1.2. Planteamiento del Problema

El ser humano posee la habilidad de analizar e interpretar datos visuales de forma rápida y sencilla, al mismo tiempo puede lidiar con una serie de dificultades impuestas por el ambiente (*e.g.*, oclusión, variación en las condiciones de iluminación, diferentes posturas del cuerpo, diferentes ángulos de visualización). Por otro lado, lo mismo no se aplica a los algoritmos de computadores, los cuales están condicionados a una serie de restricciones (*e.g.*, tiempo de procesamiento, complejidad, tasa de reconocimiento). De este modo, surge el siguiente cuestionamiento: ¿Cómo desarrollar programas de computador capaces de realizar esas tareas con eficiencia y eficacia?

Los métodos de extracción de características están innovando esta tarea y alcanzan un desempeño notable. Sin embargo, no todas las características son útiles para otros conjuntos de datos,lo que hace que el reconocimiento exacto de acciones humanas aún sea una tarea compleja, debido a obstaculos como el fondo no uniforme de la escenas y las variaciones intra-clase significativas.

Algunos de esos problemas pueden ser contornados usando datos de profundidad, ya que ellos proveen mucha información extra, que genera nuevas perspectivas para los investigadores que buscan resolver varios problemas tradicionales en visión por computador. Mientras tanto, el audio, otra modalidad muy importante en los videos, también puede proveer evidencias útiles sobre las escenas de video. Cuando cualquier información de audio o visual por si sola no es suficiente, la combinación de las mismas pueden resolver ambigüedades y ayudar a obtener respuestas mas precisas.

Al contrario de la mayoría de los métodos tradicionales que analizan los datos de forma separada, se propone la integración de la información visual y de audio para el análisis de escenas que contienen acciones humanas. El uso conjunto de informacion visual y de audio puede ayudar a extraer informacion que mejorara los resultados de reconocimiento. Por lo tanto, surge la necesidad de realizar una investigación más profunda sobre las diversas técnicas de extracción de características a partir de datos multimodales, con la intención de mejorar el poder de discriminación de los algoritmos existentes de visión por computador.

1.3. Objetivos

1.3.1. Objetivo Principal

Proponer un modelo de reconocimiento de acciones humanas con la combinación de información multimodal (intensidad, profundidad y audio).

1.3.2. Objetivos Especificos

- Realizar un estudio sobre descriptores RGB-D, en los distintos canales hasta llegar a uno que se ajuste más al canal de información indicado.
- Realizar una investigación sobre las técnicas de aprendizaje de máquina para la etapa de clasificación .
- Analizar la propuesta con distintas bases de vídeos y realizar una propia.

1.4. Contribución

Se plantea una nuevo método para el reconocimiento de acciones humanas cotidianas, el cual aporta las siguientes contribuciones:

- Utiliza información multimodal para el reconocimiento de acciones, ya que en la actualidad debido al bajo costo de la tecnología es muy común encontrar este tipo de información en cualquier dispositivo.
- Se detecta varios tipos de acciones, se debe recordar que las acciones cotidianas varían respecto al ambiente, se logra reconocer las acciones en diferentes ambientes.
- Se desarrolló una nueva base de vídeos con información multimodal para la prueba del método propuesto y para el uso de investigaciones similares.
- Se supera el promedio de asertividad en la base de vídeos HMA esto se logra dividiendo los vídeos en clips.

1.5. Organización de tesis

Se describe la organización de la presente tesis para un mejor entendimiento, la cual esta dividida en 6 capítulos, siendo este el introductorio donde es citado el problema de investigación y los objetivos que persigue esta tesis . En el Capítulo 2 correspondiente al marco teórico, se definen conceptos básicos para entender mejor la descripción de la tesis. El Capítulo 3 trata sobre el estado del arte donde se menciona trabajos con el mismo fin de reconocer acciones y el desarrollo de estos. El Capítulo 4 describe el método propuesto y sus respectivas etapas. El Capítulo 5 describe, los experimentos que se realizaron con las distintas bases de vídeos y los resultados de estas. Finalmente el Capítulo 6 muestra las conclusiones que se pudieron obtener a partir de la investigación y los trabajos futuros.

Capítulo 2

Trabajos Relacionados

El reconocimiento de acciones es importante dentro del área de visión por computador. Sus aplicaciones involucran interacciones entre personas y dispositivos tales como interfaces hombre-máquina. La mayoría de estas aplicaciones requieren un reconocimiento automático de las actividades de alto nivel.

Se han dado ya varios aportes, los que han logrado dar nuevos avances no solo en el área de visión por ordenador si no también en áreas como: procesamiento digital de señales y reconocimiento de patrones. Por esto, cada vez las aplicaciones son mas especificas, como por ejemplo : Sistemas de Protección de Personas mayores y/o niños, Fisioterapia Asistida por Ordenador, Ordenadores que hacen la labor de entrenadores en gimnasios, Análisis Semántico de Movimientos, Robótica, *etc.*

La diversidad y complejidad de los movimientos humanos llevó a buscar mejores soluciones para que no se restrinjan a movimientos específicos; por lo que se considera la misma división de investigaciones que (Herath et al., 2016).

- **Datos Visuales:**

  Las primeras investigaciones en el reconocimiento de acciones hacen uso de modelos 3D. Un ejemplo es el uso de cilindros conectados para modelar conexiones de las extremidades para el reconocimiento de peatones(Rohr, 1994). Sin embargo, la captura de los modelos 3D es muy difícil y costosa. Es por ello que las investigaciones recientes evitan el modelado en 3D y en su lugar optan por representar acciones a nivel global o local. Por lo tanto se considera la siguiente división :

  - Representaciones Holísticas: El reconocimiento de acción se basa en la extracción de una representación global de la estructura del cuerpo humano, la forma y los movimientos.
  - Representaciones Locales: El reconocimiento de acción se basa en la extracción de características locales.

- **Datos de Profundidad:**

  En contraste con imágenes de intensidad, ha habido tecnologías de visión como el sensor de profundidad que pueden capturar información de distancia del mundo real, la cual no se puede obtener de una imagen de intensidad. Donde el valor de cada píxel indica distancia entre la cámara y la escena. Las imágenes resultantes se denominan imágenes de rango o mapas de profundidad. Una de las ventajas de estos sensores es que en cada pixel tiene información de profundidad por lo que la segmentación del objeto es mucho mas sencilla y son mucho menos afectados por los cambios de iluminación.

- **Basados en Audio:**

  Se ha demostrado en varios estudios experimentales que la integración de la información de audio y de vídeo lleva a una mejora del rendimiento de reconocimiento de acciones. La información en un canal no siempre puede proporcionar evidencia acerca de las acciones, por lo que se opta tomar una percepción auditiva, ya que muchas actividades humanas producen sonidos muy característicos lo que infiere de manera efectiva las acciones humanas correspondientes. Las técnicas actuales basadas en información de audio superan condiciones como el desorden, las variaciones en las condiciones de iluminación y oclusiones totales y parciales.

2.1. Datos Visuales

La diversidad y complejidad de los movimientos humanos llevó a buscar mejores soluciones, para que no se restrinjan a movimientos específicos con el fin de superar estos problemas, un artículo reciente (Brun et al., 2014) propone Hack, un método para el reconocimiento de acciones Humanas, la idea principal de este es representar cada acción a través de una secuencia de caracteres visuales, es decir una cadena, correspondiente a las acciones elementales, construidas de acuerdo con un diccionario adquirido durante la etapa de aprendizaje, La similitud entre las acciones se evalúa con un rápido kernel de alineación global, lo que permite hacer frente a las acciones de diferente longitud.

Yilmaz y Shah identifican acciones a través de las propiedades diferenciales del espacio tiempo y volumen (STV del inglés *SpaceTime Volume*) (Yilmaz y Shah, 2005). Un STV se construye apilando los contornos de objetos a lo largo del eje del tiempo. Los cambios de dirección, la velocidad y la forma de un STV intrínsecamente caracterizan la acción subyacente. Una acción es un conjunto de propiedades extraídas de la superficie de un STV (por ejemplo, la curvatura de Gauss)

Representaciones holísticas han sobresalido en la investigación de reconocimiento de acciones debido a que estas son propensas a conservar la estructura espacial y temporal de las acciones. Sin embargo, hoy en día las representaciones locales han tomado su lugar, varias razones se atribuyen a este cambio. Por ejemplo, (Dollár et al., 2005) afirma que los enfoques holísticos son demasiado rígidos para capturar las variaciones

posibles de una acción (por ejemplo, punto de vista, la apariencia, oclusiones).

2.1.1. Representaciones Locales

Varios enfoques sobre el reconocimiento de acciones se han presentado en el pasado y la mayor parte del trabajo se ha basado en los datos de intensidad, así como se describe en (Weinland et al., 2011; Poppe, 2010); donde hacen referencia a varias investigaciones, llegando a la conclusion que la extracción de características para el seguimiento de personas es uno de los principales retos. Esto es debido a que el movimiento de personas en secuencias de vídeo involucra principalmente variaciones de escala y traslación.

Hay métodos muy eficientes como (Huang y Leng, 2010; Hu et al., 2007) los cuales son muy utilizados ya que son robustos ante variaciones de escala, rotación y traslación, donde (Hu et al., 2007) demuestra que el uso de Wavelet momentos invariantes y Wavelet redes neuronales puede lograr una mayor exactitud de la clasificación de imágenes que el algoritmo basado en momentos invariantes normales y las redes neuronales BP. En algunos casos la cantidad de características son de altas dimensiones, lo que ocasiona el uso de otros algoritmos, Yan et al. (2011) resume los recientes trabajos de investigación relacionados con la idea de descomponer los patrones de altas dimensiónes en patrones de baja dimensionalidad y enfoques para lograr el óptimo global en relación con la asertividad y la complejidad del tiempo

Puntos de interés espacio-temporales como se propone en (Laptev, 2005) donde se captura el volumen en la variación de intensidad de gris de los pixeles tanto en el dominio del espacio y del tiempo , esta bidimensionalidad es utilizada en varias investigaciones dando muy buenos resultados,incluyendo la metodología propuesta.

2.1.1.1. Detección de puntos de interés:

Laptev extiende el detector de esquinas Harris (Harris y Stephens, 1988). En Harris-3D, se obtiene estructuras espaciales robustas. La idea del detector de esquinas 2D Harris es encontrar localizaciones espaciales de una imagen con cambios significativos en dos direcciones ortogonales. El detector Harris-3D identifica puntos con grandes variaciones espaciales y los movimientos no constantes(Laptev, 2005)

Otro detector de punto de interés 2D ampliamente utilizado, es el detector Hessian, se extiende también a su contraparte en 3D (Willems et al., 2008). A diferencia del detector de 3D-Harris, donde los gradientes se utilizan para la detección de puntos de interés, este hace uso de las segundas derivadas.

Un mejor enfoque para el reconocimiento de acciones fue el tomar en cuenta características locales espacio-temporales, estas se han investigado ampliamente sobre

el canal de intensidad. El enfoque más popular es la de representar una acción humana usando partes del cuerpo articulados como lo hacen en(Sheikh et al., 2005; Yilma y Shah, 2005) donde (Sheikh et al., 2005) basa su trabajo en una pantalla de punto de luz como representación de la postura a través de un conjunto de puntos en el espacio, consideran que las tres fuentes más importantes de la variabilidad en la tarea de reconocer las acciones provienen de las variaciones de: punto de vista, la tasa de ejecución, y la antropometría de los actores.

2.1.1.2. Descriptores Locales:

Se debe tener en cuenta la definición de un paralelepípedo 3D o simplemente un paralelepípedo, este es un cubo construido a partir de píxeles detectados alrededor de los puntos de interés, así lo considera (Dollár et al., 2005; Laptev, 2005). Investigaciones distintas como (Messing et al., 2009) cuestiona la elección del paralelepípedo e introducen la noción de trayectorias,se debe considerar que los descriptores locales se pueden emplear en paralelepípedos y trayectorias.

- **Descriptor de esquinas y movimientos:** (Klaser et al., 2008) sugiere usar el histograma de orientacion de gradientes como un descriptor de movimiento. Aunque inspirado por la robustez de este en el reconocimiento de imágenes (Dalal y Triggs, 2005) extendieron el dominio espacio-temporal, por lo tanto, renombraron el descriptor como HoG3D.
- **Descriptor de patrones binarios:** Patrones binarios locales son descriptores 2D basados en intensidad, que se utilizan con éxito en una amplia gama de problemas, incluyendo el reconocimiento de rostros y análisis de la textura como en (Ojala et al., 2002) donde basa el reconocimiento en patrones binarios locales a los que denomina 'uniformes' estos son propiedades fundamentales de la textura de la imagen local y su histograma de ocurrencia resulta ser una función muy potente para la textura. El descriptor de patrones se calcula mediante la cuantificación de la vecindad de un píxel con respecto a su intensidad.

2.2. Datos de Profundidad

La metodología propuesta hace uso de mapas de profundidad, obtenidos de un sensor RGB-D (Kinect), como lo han hecho investigaciones tales como (Broggi et al., 2000) el cual detecta peatones usando este canal de información , como también lo hicieron. (López et al., 2014) Obtiene siluetas de gran calidad, las cuales pueden ser obtenidas incluso en total carencia de iluminación ya que este canal de información da una gran ventaja respecto a la segmentación.

En (Li et al., 2010), los resultados experimentales demuestran mas del 90 % de precisión de reconocimiento y este se logra mediante el muestreo en puntos 3D a partir

de mapas de profundidad. En comparación con el reconocimiento basado en siluetas 2D, se reducen a la mitad los errores de reconocimiento. Además, se demuestra el potencial de la bolsa de puntos para hacer frente a las oclusiones. Otros trabajos de reconocimiento de acciones basados en profundidad que obtuvieron buenos resultados son (Junejo et al., 2011; Gilbert et al., 2008) , (Gilbert et al., 2008) se basa en buscar esquinas espacio-temporales y determinar la disposición espacial relativa de todas las demás esquinas en el marco y (Junejo et al., 2011) aborda el reconocimiento de las acciones humanas en virtud de los cambios de vista, explora auto-similitud de secuencias de acción en el tiempo y observar la estabilidad notable de este tipo de medidas a través de puntos de vista, basándose en esta observación clave, desarrolla un descriptor de acción que captura la estructura de las similitudes y diferencias temporales dentro de una secuencia de acción.

La aparición del sensor de profundidad a bajo costo (Kinect) ha acrecentado las investigaciones en visión por ordenador, juegos de azar, reconocimiento de gestos, y la realidad virtual. (Shotton et al., 2013) propone un método para predecir las posiciones de las articulaciones del cuerpo en 3D, tambien se propone reconocer un accion mediante una imagen simple de Kinect ,(Xia et al., 2012) propuso un modelo basado en el algoritmo para detectar humanos usando mapas de profundidad generados por Kinect. (Li et al., 2010) emplea un gráfico de acción para modelar la dinámica de las acciones y muestra una bolsa de puntos 3D a partir del mapa de profundidad para caracterizar un conjunto de posturas.

La escena capturada de una cámara de profundidad puede combinarse con posiciones 3D, entonces junto con la posición 3D de una persona se reconoce la actividad (Jansen et al., 2007) utiliza una restricción de distancia sencilla, utiliza la altura de la silueta de la persona a reconocer, así puede diferenciar si la persona está de pie, sentado o acostado. Donde afirman que este reconocimiento es útil para el cuidado de un hogar de ancianos. (Chen et al., 2011), con el mismo objetivo de reconocer actividades domésticas tales como beber, utiliza la distancia entre las partes del cuerpo y los objetos a través del tiempo, y los modelos de cada actividad a través de razonamiento espacio-temporal mediante el intervalo de Álgebra de Allen.

Dynamic Time Warping (DTW) es otra técnica tambien utilizada (Sempena et al., 2011) lo usa debido a su gran robustez frente a la variación de la velocidad o el estilo en la realización de una acción, para mejorar la tasa de acierto realiza el seguimiento de partes del cuerpo mediante el uso de cámara de profundidad, creando su propio vector de características.

Obteniendo también muy buenos resultados pero con un nuevo método para el reconocimiento de la acción humana con histogramas de ubicaciones conjuntas 3D (HOJ3D) está (Xia et al., 2012) que usa coordenadas esféricas modificadas y HMM se aplican a la tarea de clasificación, la ventaja principal es el rendimiento en tiempo real.

.

La combinación multimodal RGBD ha dado buenos resultados , pero aun no son óptimos para las investigaciones por lo que en la presente propuesta tomaremos un canal mas de información .

2.3. Datos de Audio

Los datos visuales y de profundidad proporcionan información robusta de la escena y permiten que se pueda reconocer acciones incluso de una sola imagen. Recientemente, datos de la gama 3D o RGB-D también se han hecho populares para el reconocimiento de acciones, por lo que plantean reconocer una gran clase de las actividades humanas. Sin embargo, estas modalidades se limitan al campo de visión de la imagen y no es robusto en todos los rangos de condiciones ambientales.

Por otra parte, la información no siempre puede proporcionar evidencia acerca de las acciones de un ser humano, el enfoque que se opta tomar es la percepción auditiva, ya que muchas actividades humanas producen sonidos muy característicos de la que puede inferir de manera efectiva las acciones humanas correspondientes.

Existen varias investigaciones en el campo de la extracción de características basadas en audio, coeficientes ceptrales de mel es una de las herramientas mas utilizadas ya que es muy robusta, (Hasan et al., 2004) presenta un sistema de seguridad basado en la identificación del hablante donde utiliza este método junto con la cuantificación vectorial para minimizar la cantidad de datos a manejar.

A pesar de que MFCC esté diseñado para la tarea de reconocimiento de voz , se ha utilizado para describir un gran número de diferentes clases de sonido como lo hicieron (Breebaart y McKinney, 2004; Chen et al., 2005; Eronen et al., 2006), donde (Breebaart y McKinney, 2004) detecta distintos sonidos como el ruido, la música y el silencio; (Chen et al., 2005) se enfoca en sonidos característicos que ocurren en un baño donde utiliza Hidden Markov Model(HMM) y MFCC, (Eronen et al., 2006) también hace uso de HMM y considera categorías de sonido como las secuencias de pequeñas muestras de estas,obteniendo 88 % de asertividad,es por eso que también se tomará en cuenta en este trabajo.

Trabajos anteriores se basan en la segmentación del flujo de audio para reconocer las actividades humanas (Harma et al., 2005; Zhang y Kuo, 2001; Breebaart y McKinney, 2004; Peltonen et al., 2002) usualmente lo han logrado a través de la detección de silencio, la detección de cambios en las características abruptas o incluso anotación manual. Otros métodos que no se basan en la segmentación o en hacer uso de procesamiento por lotes son (Eronen et al., 2006; Zhu et al., 2007) asumiendo un mínimo de tiempo de duración de las actividades, o clasificar sólo las cracterísticas de audio de corta duración.

El enfoque presentado por(Wang et al., 2003) consiste en una técnica de reconocimiento rápido para una gran base de datos de canciones, es capaz de generalizar

varias categorías de sonidos y gestionar múltiples errores de clasificación.

En la literatura contamos con investigaciones sobre el reconocimiento de acciones basadas en audio (Stork et al., 2012) usando HMM llega a reconocer 22 diferentes acciones con este canal;la información multimodalidad tiene hasta ahora relativamente poca atención en áreas de investigación, investigaciones muy recientes como (Pieropan et al., 2014a) donde utilizan información multimodal obtiene una acertividad del 73 % donde utilizan MFCC para el canal de audio.

Para la **clasificación** la técnica de Bolsa-de-palabras (BOW, del ingles *Bag of Words*) se ha aplicado popularmente en enfoques de características locales espacio-temporales, en un enfoque tradicional BOW, los descriptores de características se agrupan en palabras visuales, y la acción se convierte en un solo histograma de la frecuencia de ocurrencias de palabras visuales. Este histograma representa el vídeo en una forma compacta, y es robusto al espacio y el tiempo como lo hacen (Schuldt et al., 2004; Laptev et al., 2008a; Bilinski y Bremond, 2011).

2.4. Consideraciones Finales

Varias investigaciones en la literatura actual han sido hechas utilizando un único canal de información. En los últimos años, nuevo dispositivos han aparecido, siendo capaces de capturar otras fuentes de infomación como intensidad, profundidad y audio. Esta propuesta busca utilizar todas esas fuentes de información para conseguir detectar acciones en videos.

Capítulo 3

Marco Teórico

El presente capítulo cuenta con definiciones y explicaciones breves, agrupadas según categorías para brindar mayor claridad sobre los conceptos presentes en esta tesis.

3.1. Conceptos Básicos

3.1.1. Imagen

Una imagen es una función bidimensional de intensidad de la luz $f(x, y)$ donde x e y denotan las coordenadas espaciales, el valor de f en cualquier punto (x, y) es proporcional al brillo de la imagen en ese punto.Una imagen digital es una imagen $f(x, y)$ discretizada tanto en coordenadas espaciales como en brillo. Es una matriz de puntos llamados **Pixeles** (un pixel representa a la menor unidad homogénea en una imagen), una imagen a color está formada por tres canales RGB que corresponden al rojo(RED), verde ($GREEN$) y azul ($BLUE$) como se observa en la Figura 3.1.

3.1.2. Vídeo

Un vídeo es una sucesión de imágenes presentadas a cierta frecuencia. La fluidez de un vídeo se caracteriza por el número de imágenes por segundo (frecuencia de cuadros), expresado en **FPS** (*cuadros por segundo*) como se muestra en la Figura 3.2.

3.1.3. Canal de Intensidad

La intensidad es la cantidad de luz emitida por un punto en la escena, por lo general se estima como el promedio de los tres canales y para mantener los valores

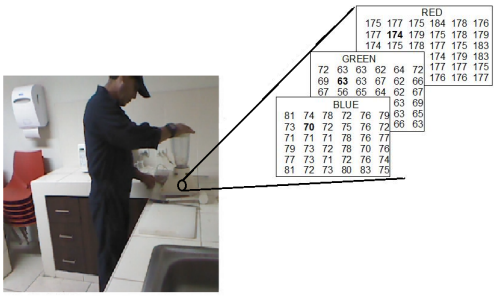


Figura 3.1: Cada pixel va a contar con tres valores numéricos en un rango de 0 a 255 (0 indica la ausencia de un color y 255 la máxima representación de ese color en ese punto.)

Figura 3.2: El ojo humano es capaz de distinguir aproximadamente 20 imágenes por segundo. De este modo, cuando se muestran más de 20 imágenes por segundo, es posible engañar al ojo y crear la ilusión de una imagen en movimiento (SENA, 2009) .

entre 0 a 255 se normaliza. Como se muestra en la Figura 3.3.

3.1.4. Canal de Profundidad

Los sensores de profundidad como el Kinect u otros utilizan las posiciones relativas de los puntos en el patrón para calcular el desplazamiento de profundidad en cada posición de pixel en la imagen. Cabe señalar que los valores de profundidad reales son la distancia desde el plano de la cámara láser así como se observa en la Figura 3.5.Los sensores también brindan mapas de profundidad, donde se muestra con un tono oscuros (en esta caso azul) la cercanía a la cámara y con un tonos cálidos (en este caso rojo) lo mas lejano a la cámara como en la Figura 3.4.

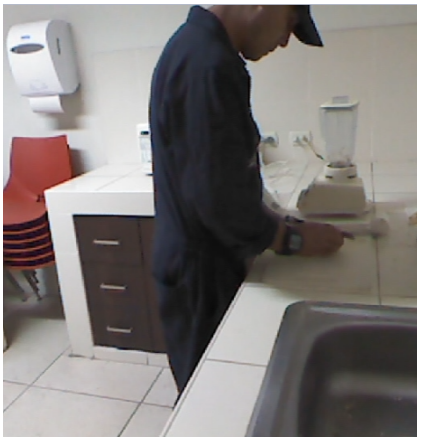


Figura 3.3: Canal intensidad.

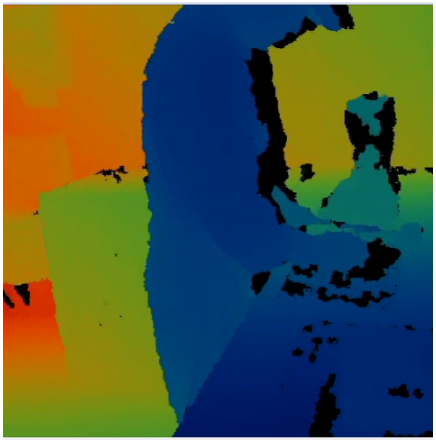


Figura 3.4: Canal Profundidad.

3.2. Descriptores

Los descriptores son representados en vectores donde como su nombre lo indica, describen las características de los contenidos dispuestos en imágenes o en vídeos. Describen características elementales tales como la forma, el color, la textura o el movimiento, entre otros. Estos toman valores en los reales, en los que cada dimensión recibe un significado según el tipo de característica que se esté midiendo.

3.2.1. Histogramas Orientados de Gradientes

HOG (del ingles *Histogram of Oriented Gradients*)presentado por (Dalal y Triggs, 2005), es un descriptor de características de forma utilizados para la detección de objetos en visión por computador y procesamiento de imágenes.

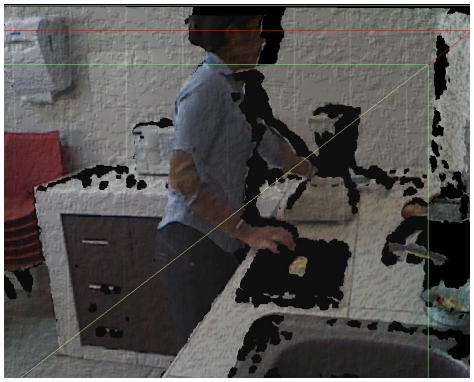


Figura 3.5: Canal Profundidad, imagen generada por Kinect 3D Viewer.

En términos generales, la imagen es dividida en un conjunto de celdas uniformes. El algoritmo estima la orientación del gradiente de los píxeles que conforman cada celda como se observa en la Figura 3.6 y reúne la información en un histograma de orientaciones de N clases. Para mayor rendimiento el algoritmo realiza una normalización de contraste.

Figura 3.6: En la parte izquierda se observa la imagen de entrada, en la parte derecha se observa las magnitudes de las imágenes, que al final terminan siendo una representación numérica.

En el proceso de extracción de características, se divide la imagen en celdas (una celda consiste en una región de la imagen que mide 8 pixeles de alto por 8 pixeles de ancho). Luego, por cada pixel dentro de una celda se acumulan los histogramas de orientación de los gradientes. Estos histogramas capturan propiedades (bordes) de forma local dentro de la celda. A su vez, estos histogramas son invariantes a pequeñas deformaciones.

El gradiente en cada pixel está discretizado en uno de los nueve contenedores de orientación, se debe tomar en cuenta que un histograma está compuesto por contenedores. Un contenedor es la representación de la cuantificacion de un espacio. En

el contenedor se cuenta la cantidad de elementos del espacio que existen en la región definida por dicho contenedor.

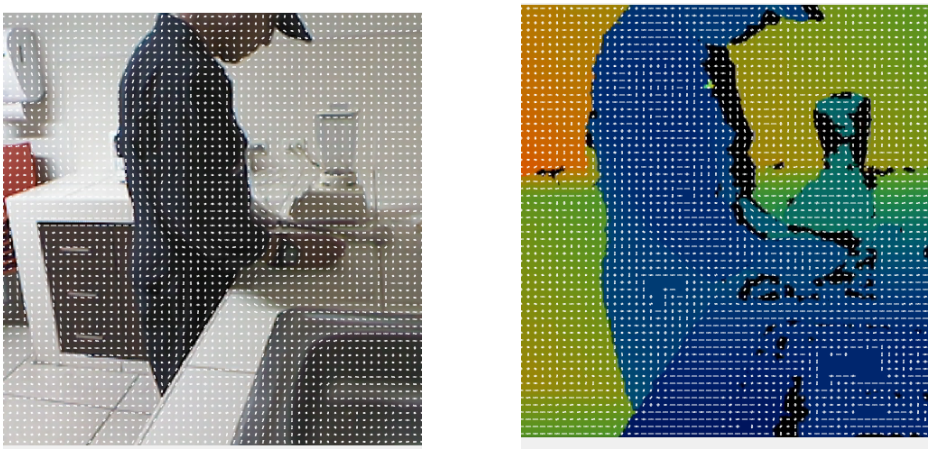


Figura 3.7: En las imágenes tanto de la izquierda (intensidad) como derecha(profundidad) se muestra las magnitudes por celda

Para las imágenes en color, se calcula el gradiente por cada canal, un canal por color eligiendo el canal que posea una mayor cantidad de pixeles con magnitudes elevadas.

Finalmente, cada histograma de las celdas es normalizado con respecto a la energía del gradiente en un vecindario alrededor de la celda. La energía del gradiente se obtiene multiplicando cada posición del vector de características por un factor definido, esa energía se calcula como:

$$N_{\delta},\gamma(i,j) = (C(i,j)^2 + C(i+\delta,j)^2 + C(i,j+\gamma)^2 + C(i+\delta,j+\gamma)^2)^{\frac{1}{2}} \tag{3.1}$$

donde, $\delta, \gamma \in \{-1, 1\}$.Luego se normaliza el histograma por cada celda dada con respecto a la energía total en cada uno de estos bloques. Esto entrega un vector de longitud 9 x 4 que representa la información local del gradiente dentro de una celda.

3.2.2. Puntos de Interés Espacio-Temporales

Los Puntos de Interés Espacio-Temporales (STIP, del ingles *Space-Time Interest Points*) es un algoritmo desarrollado por (Laptev et al., 2008a) que utiliza la técnica de los puntos de interés a lo largo del espacio y tiempo; Estos puntos de interés consisten

en una evolución de la idea original de Harris para detectar puntos de interés en un dominio espacial.

Dichos puntos de interés originales consisten en un detector que busca puntos en la imagen donde existan cambios significativos en ambas direcciones, tanto en horizontal como en vertical. Un concepto importante en visión por computador es la escala de observación σ^2, el cual viene a expresar la granularidad del detalle con la que se trata una imagen. Un ejemplo sencillo para su entendimiento es pensar en una arboleda. Si la escala de observación es de unos pocos centímetros podremos diferenciar las ramas y las hojas de los árboles; mientras que si la escala es de cerca de medio metro, podremos distinguir los troncos de los árboles pero no sus ramas, quedando éstas en un efecto difuminado.

Para una escala de observación dada σ_ι^2 , los puntos de interés con cambios en la componente vertical y horizontal pueden ser detectados usando el segundo momento matricial integrado dentro de una ventana Gaussiana con varianza σ_i^2. Para ello se aplicará a cada uno de los puntos de la imagen el operador definido en la Ecuación 2.1

$$\begin{aligned} \mu^{sp}(\cdot;\sigma_\iota^2;\sigma_i^2) &= g^{sp}(\cdot;\sigma_i^2) * ((\bigtriangledown L(\cdot;\sigma_\iota^2))(\bigtriangledown L(\cdot;\sigma_\iota^2))^T) \\ &= g^{sp}(\cdot;\sigma_i^2) * \begin{pmatrix} (L_x^{sp})^2 & L_x^{sp}L_y^{sp} \\ L_x^{sp}L_y^{sp} & (L_x^{sp})^2 \end{pmatrix} \end{aligned} \tag{3.2}$$

donde '*' denota el operador de convolución, $L^{sp}(x,y;\sigma_\iota^2)$ es la imagen $f^{sp}(x,y)$ en su representación lineal a escala σ_ι^2

$$L^{sp}(x,y;\sigma_\iota^2) = g^{sp}(x,y;\sigma_\iota^2) * f^{sp}(x,y)$$

y L_x^{sp} y L_y^{sp} son sus derivadas en la escala σ_ι^2.

$$\begin{aligned} L_x^{sp} &= \delta_x(g^{sp}(\cdot;\sigma^2\iota) * f^{sp}(x,y)) \\ L_y^{sp} &= \delta_y(g^{sp}(\cdot;\sigma^2\iota) * f^{sp}(x,y)). \end{aligned}$$

Siendo λ_1 y λ_2 ($\lambda_1 \leq \lambda_2$) los valores propios de μ^{sp}; dos valores significativamente altos de λ_1 y λ_2 indican la presencia de un punto de interés. Para detectarlos, Harris y Stehens propusieron encontrar máximos positivos en la Ecuación 3.3.

$$\begin{aligned} H^{sp} &= det(\mu^{sp}) - ktrace^2(\mu^{sp}) \\ &= \lambda_1\lambda_2 - k(\lambda_1+\lambda_2)^2 \end{aligned} \tag{3.3}$$

Donde exista un punto de interés, el ratio de los valores propios $\alpha = \lambda_2/\lambda_1$ debe de ser alto. Es decir, para los máximos locales positivos de H^{sp}, el ratio α tiene que satisfacer $k \leq \alpha/(1+\alpha)^2$. Un valor de k comúnmente utilizado en la bibliografía es k = 0.004 lo que corresponde a la detección de puntos con $\alpha < 23$.

Partiendo de esta base, Laptev et al. (2008a) extendió la noción de punto de interés añadiendo la componente temporal; requiriendo que los puntos tengan grandes variaciones en la componente espacial y temporal. Los puntos resultantes corresponden a puntos de interés espaciales en distintos momentos del tiempo donde se realiza un movimiento no constante. Para ello desarrolló un operador, definido en la Ecuación 3.4 que responde a eventos en secuencias de imágenes temporales en una determinada localización y con una extension específica en el espacio y en el tiempo. Al igual que en el dominio espacial, se utiliza el segundo momento matricial usando una función Gaussiana de ponderación $g(\cdot; \sigma_i^2; \tau_i^2)$

$$\mu = g(\cdot; \sigma_i^2; \tau_i^2) * \begin{pmatrix} L_x^2 & L_x L_y & L_x L_t \\ L_x L_y & L_y^2 & L_y L_t \\ L_x L_t & L_y L_y & L_t^2 \end{pmatrix} \tag{3.4}$$

donde L es la representación espacio temporal de una secuencia de imágenes

$$L(\cdot; \sigma_\iota^2; \tau_\iota^2) = g(\cdot; \sigma_\iota^2; \tau_\iota^2) * f(\cdot)$$

y L_x, L_y y L_t son sus derivadas en las escalas σ_ι^2 y τ_ι^2:

$$L_x(\cdot; \sigma_\iota^2; \tau_\iota^2) = \delta_x(g * f)$$

$$L_y(\cdot; \sigma_\iota^2; \tau_\iota^2) = \delta_y(g * f)$$

$$L_t(\cdot; \sigma_\iota^2; \tau_\iota^2) = \delta_t(g * f).$$

Para detectar los puntos de interés, se buscan regiones en f que tengan valores significativos en los valores propios λ_1, λ_2 y λ_3 de μ. Para ello se modificó la función de Harris definida en el dominio espacial(3.3), añadiendole el dominio temporal, obteniendo la expresión de la Ecuación 3.5. Este detector de puntos de interés se conoce como Harris3D por su sustento en el dectector espacial original de Harris.

$$\begin{aligned} H &= \det(\mu) - k\ trace^3(\mu) \\ &= \lambda_1\lambda_2\lambda_3 - k(\lambda_1 + \lambda_2 + \lambda_3)^3. \end{aligned} \tag{3.5}$$

Los puntos de interés espacio temporales de f pueden ser encontrados detectando los máximos locales positivos de H. Definiendo los ratios $\alpha = \lambda_2/\lambda_1$ y $\beta = \lambda_3/\lambda_1$ se puede reescribir la Ecuación 3.5 como:

$$H^{sp} = \lambda_l^3(\alpha\beta - k(1+\alpha+\beta)^3)$$

y en los maximos locales positivos de H obtenemos $k \leq \alpha\beta/(1+\alpha+\beta)^3$, siendo su máximo valor $k = 1/27$ cuando $\alpha = \beta = 1$. Esto indica que para valores suficientemente altos de k, los máximos locales positivos de H corresponden a puntos con grandes variaciones a lo largo de la componente espacial y de la temporal. En particular, si fijamos el valor de α y β a 23 como en el caso espacial, obtenemos que el valor de k utilizado en H es $k \approx 0{,}005$

El algoritmo STIP brinda la opción stipshow el cual muestra los puntos de interés constantes a través de los frames como muestra la Figura 3.8.

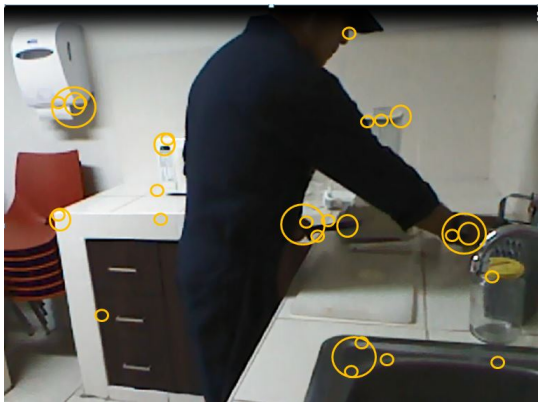


Figura 3.8: Las circunferencias tienden a crecer cuando es continuamente constante en los frames

3.2.3. Mel Frequency Cepstral Coefficients

MFCC (del ingles *Mel Frequency Cepstral Coefficients*)Los MFCC se utilizan para el reconocimiento de señales de voz, se trata de una serie de filtros triangulares, cuyas frecuencias centrales están separadas en base a una escala de Mel, es decir, que el incremento en frecuencia que percibe el oyente es directamente proporcional al incremento de valor dentro de esta escala. Después de filtrar con los filtros de Mel, se obtiene una serie de coeficientes que indican la energía de cada banda a la salida del filtro.

La implementación estándar del cálculo del MFCC(Niemann, 2013) se muestra en la Figura 3.9 a continuación se explicará cada uno de los pasos:

3.2.3.1. Transformada de Fourier

La primera etapa de tratamiento es el cálculo de la representación en el dominio de frecuencia de la señal de entrada. Esto se consigue mediante el cálculo de la transformada de Fourier discreta como se muestra en la Ecuación 3.6

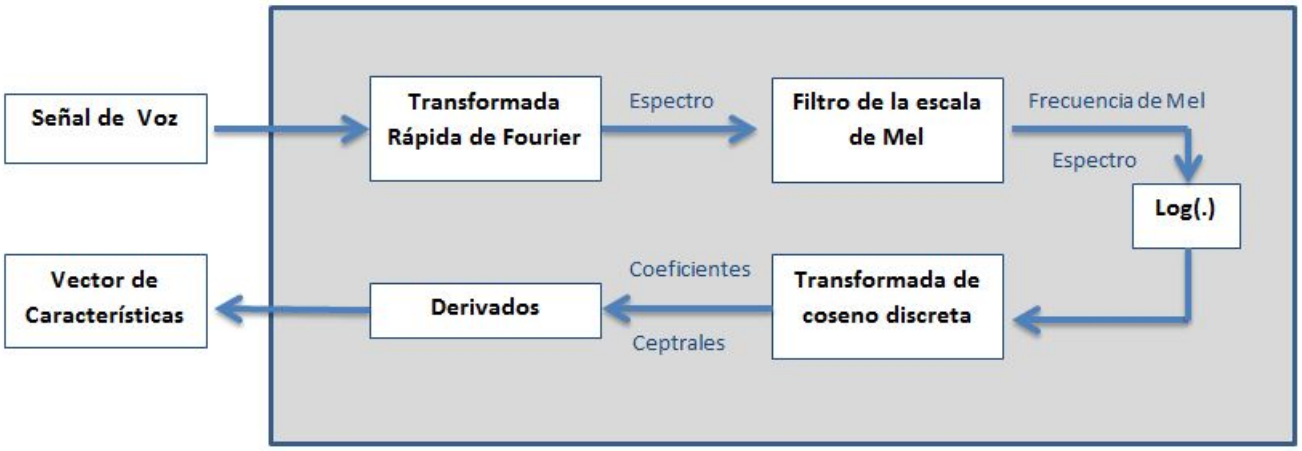


Figura 3.9: Diagrama del algoritmo de MFCC

$$c_{\tau,j}^{(3)} = \log(c_{\tau,j}^{(2)}) \quad \text{donde j= } 0, 1, ...N_d \tag{3.6}$$

Donde N es el número de puntos de muestreo dentro de una trama de voz y τ el marco de tiempo. Para implementaciones, la transformada de Fourier rápida, que es una variación de la Transformación Discreta de Fourier optimizado para la velocidad(Niemann, 2013).

3.2.3.2. Filtro de la escala de Mel

El segundo paso de procesamiento es el cálculo del espectro de frecuencia de Mel. Por lo tanto, el espectro es filtrado con N_d distintos filtros de paso de banda así se procede a calcular el poder de cada banda de frecuencia. Este filtrado imita el oído humano porque este sistema utiliza la potencia en una banda de frecuencia de la señal para su posterior procesamiento. Esta etapa de procesamiento puede ser descrita por la Ecuación 3.7:

$$c_{\tau,j}^{(2)} = \Sigma_{k=0}^{N/2-1} d_{j,k} C_{\tau,j}^{(1)} \quad \text{donde j= } 0, 1, ...N_d \tag{3.7}$$

Donde d es la amplitud del filtro de paso de banda con el índice j a la frecuencia k.
La escala de Mel es una escala no lineal que se adapta a la percepción del tono no lineal del sistema auditivo humano. El número, la forma (triangular, rectangular trapezoidal) y la frecuencia central de los filtros de paso de banda pueden variar (Niemann, 2013). La Figura 3.10 muestra un banco de filtros típico con 25 filtros de paso de banda triangulares.

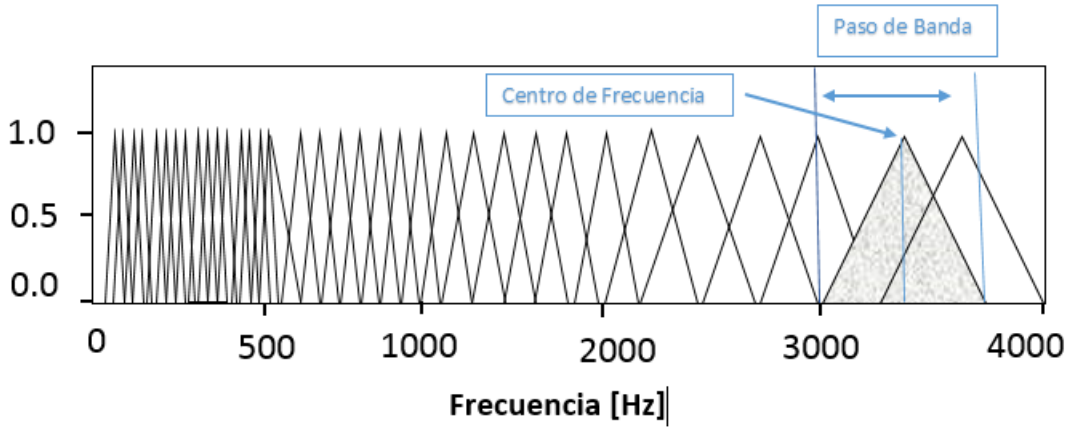


Figura 3.10: Muestra un banco de filtros típico con 25 filtros de paso de banda triangulares

3.2.3.3. Logaritmo

La tercera etapa de procesamiento calcula el logaritmo de la señal, para imitar la percepción humana de la sonoridad porque los experimentos mostraron que los seres humanos perciben la sonoridad en una escala logarítmica (Niemann, 2013).

3.2.3.4. Transformada de coseno discreta

La cuarta etapa de procesamiento trata de eliminar las características dependientes del hablante mediante el cálculo de los coeficientes cepstrum. A partir del modelo de fuente y filtro se sabe, que la señal es la convolución de la señal fuente dependiente del altavoz y la señal de filtro. Para suprimir la señal de la fuente se calcula el cepstrum. El cepstrum puede ser interpretado como el espectro de un espectro. Por lo tanto, los armónicos dependientes del hablante de la frecuencia fundamental se transforman en un coeficiente cepstral en condiciones ideales. La transformación inversa de los coeficientes cepstrales inferiores muestran la respuesta de frecuencia del tracto vocal y la transformación inversa de la orden superior coeficientes cepstrales muestran el espectro de frecuencia de la señal fuente. Por lo tanto, los armónicos dependientes del hablante se suprimen mediante la adopción de los coeficientes cepstral de orden inferior para su posterior procesamiento. El cepstrum de una señal se calcula:

$$F^{-1}\log\left(Ff_n\right) \tag{3.8}$$

Donde f es la señal de entrada y F es la transformación de Fourier. El cálculo del logaritmo se puede omitir debido a que el logaritmo de la señal se calcula en la

etapa de procesamiento anterior. En lugar de la Transformada de Fourier se puede utilizar la transformada de cosenos discreta porque el valor absoluto del espectro es respectivamente la continuación periódica de la señal(Niemann, 2013).

3.2.3.5. Derivados

Todas las etapas de procesamiento anteriores incluyen información acerca de la señal actual. Para representar la naturaleza dinámica del discurso la primera y segunda derivadas de orden de los coeficientes cepstrum amplían el vector de características. Un vector típico de característica MFCC se calcula a partir de una ventana con 512 puntos de muestra y se compone de 13 coeficientes cepstrales, 13 derivados de primer orden y 13 derivados de segundo orden. Este ejemplo podría reducir la dimensionalidad de 512 a 39 dimensiones.

3.2.4. Espectograma

Un espectro puede entenderse como una sucesión temporal de números. Las sucesiones de números reales se pueden escribir como combinaciones lineales de senos y cosenos (o exponenciales complejas).

Cada una de las tramas que se obtienen del cálculo del STFT(Transformada de Fourier de Tiempo Reducido) se indexan en una matriz; esta representa la variación en el espectro y la energía de la señal para cada una de la sucesión de tramas a lo largo del tiempo. A medida que se van obteniendo nuevas tramas, se indexan de forma consecutiva en la primera posición de la matriz, empujando la trama anterior a la segunda posición, y la segunda a la tercera, y así sucesivamente. De esta manera se puede representar la variación del espectro de la señal y la energía en función del tiempo.

Una forma de representación del espectrograma es: el tiempo en el eje de abscisas, sucesiones consecutivas de transformadas de Fourier, en el eje de ordenadas la frecuencia expresada en Hz y representada como la mitad del espectro, ya que la transformada de Fourier es periódica y su espectro se repite a lo largo del tiempo. Y por último, la representación de la energía expresada en dB como el módulo de la amplitud de la Transformada de Fourier [20*log10(abs(X(f))] y representada con una paleta de colores, o con niveles de gris, en el caso concreto en la escala de grises, con valores donde la energía es mayor representados con unos niveles más oscuros, y aquellos valores donde la energía es más pequeña con unos niveles más claros como se observa en la Figura 3.12.

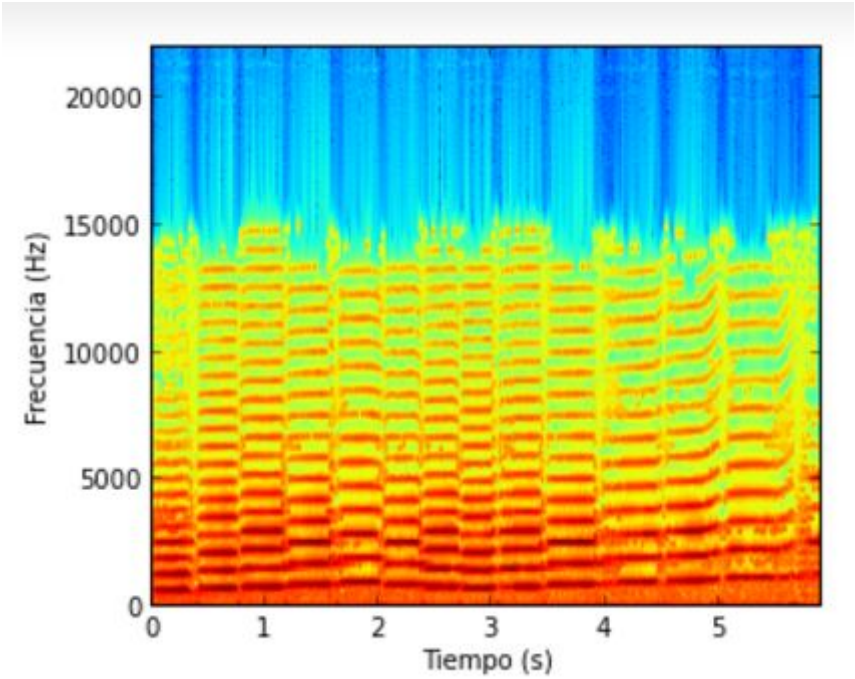


Figura 3.11: Modelo de espectograma

3.3. Aprendizaje de Máquina

El aprendizaje automático o aprendizaje de máquinas *(del inglés, Machine Learning)* es una rama de la inteligencia artificial cuyo objetivo es desarrollar técnicas que permitan a las computadoras aprender.

3.3.1. Clustering

Clustering es el proceso de agrupar datos en clases o **clusters** de tal forma que los objetos de un cluster tengan una similaridad alta entre ellos, y baja (sean muy diferentes) con objetos de otros clusters, los cuales se caracterizan por: (García y Gómez, 2012).

- Escalabilidad: normalmente corren con pocos datos.
- Clusters de formas arbitrarias: los que son basados en distancias numéricas tienden a encontrar cluster esféricos.
- Capacidad de manejar diferentes tipos de atributos: numéricos(lo mas común), binarios, nominales, ordinales, etc.
- Capacidad de añadir restricciones.
- Manejo de ruido: muchos son sensibles a datos erróneos.

- Poder funcionar eficientemente con alta dimensionalidad.
- Requerimientos mínimos para especificar parámetros, como el numero de clusters.
- Independiendentes del orden de los datos.
- Que los clusters sean interpretables y utilizables

3.3.2. Clasificadores

Se encuentran dentro del aprendizaje de máquinas y se encargan de determinar a que subconjunto corresponde un nuevo objeto, basandose en un conjunto de datos de entrenamiento, del que se conocen las clases existentes y al cual está asociado cada uno de sus objetos.

3.3.2.1. Support Vector Machine

Un clasificador SVM (del inglés *Support Vector Machine*) (Cortes y Vapnik, 1995), divide al espacio de vectores de características mediante un hiperplano, es decir, durante el proceso de entrenamiento se encarga de determinar parámetros adecuados que permitan separar los datos en dos subespacios, de tal forma que , al ingresar una consulta se calcula hacia que lado del hiperplano le corresponde ser asignada, quiere decir que su clase es aquella que esta asociada a esa zona, esto corresponde a un clasificador binario(admite dos tipos de clases), sin embargo, para extender el algoritmo de SVM a un problema como el de esta investigación, se emplea más de un modelo SVM, lo que entrega la posibilidad de realizar múltiples divisiones del espacio y filtrar la zona correcta para una nueva consulta.

3.4. Bolsa de Palabras

BOW (del ingles *Bag of words*)es un *framework* que consta de tres pasos: el primero es la extracción de características , el segundo la creación de un diccionario visual para la formación de un histograma de frecuencias y tercero la clasificación de estos histogramas mediante un entrenamiento(Sivic y Zisserman, 2003).

BoW ha demostrado tener una gran eficiencia de representación en este contexto. Este modelo fue inicialmente introducido en el procesamiento del lenguaje natural, en donde cada documento es representado por un histograma de frecuencias de palabras.

Para que esta técnica pueda ser utilizada con imágenes o videos, las mismas deben ser consideradas como documentos, *i.e.* , la imagen es considerada como una colección de eventos locales de interés, usualmente llamadas conceptos visuales. La información

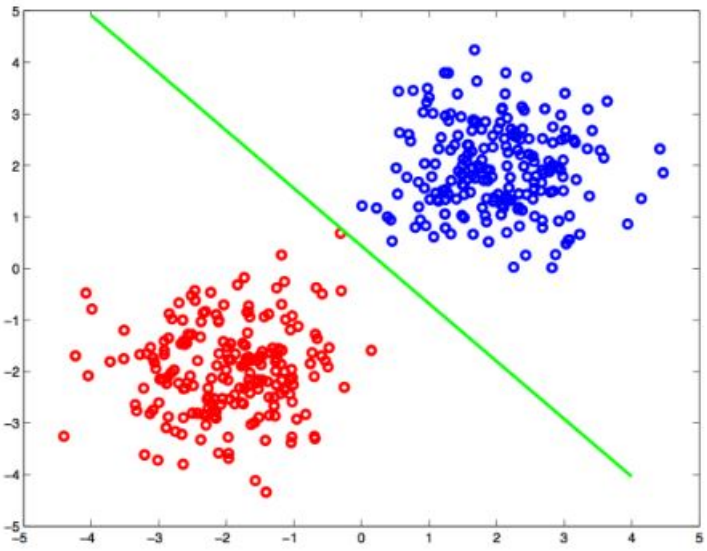


Figura 3.12: Se muestra un ejemplo sencillo de clasificación de datos de entidad: los datos dados en dos dimensiones, si los puntos rojos y puntos azules representan diferentes categorías, el problema de clasificación efectivamente se reduce a la elaboración de un límite que separa los dos conjuntos de puntos

respecto a la presencia o no de estos conceptos visuales, sirve como un indicador del contenido de la imagen.

Los conceptos visuales pueden ser generados de diversas formas, usualmente a través de la extracción de características (atributos) alrededor de puntos de interés, regiones, bordes, entre otros. Luego, esos conjuntos de datos son agrupados por técnicas de agrupamiento (*clustering*) con el objetivo de identificar los diversos grupos en el espacio de características, cada grupo es considerado como una palabra visual (*codeword*). Un conjunto de palabras visuales producen un diccionario visual (*codebook*).

Finalmente, es generado un histograma de ocurrencias de palabras visuales para cada imagen, las mismas serán utilizadas para representar la información de la imagen y que será utilizada en la etapa de clasificación.

3.4.1. Vocabulario Visual

Para poder formar el vocabulario utilizamos K-Means o K-Medias *(algoritmo de agrupamiento)*, este algoritmo tiene como objetivo la partición de un conjunto X en n elementos $X = \{x_1, x_2, ..., x_n\} \subset$ en k subconjuntos $(k \leq n)$ $H = \{H_1, H_2, ..., H_k\}$ de tal manera que se minimice el valor de J en la función:

$$J = \sum_{i=1}^{m} \sum_{j=1}^{k} r_{ij} \parallel x_i - H_j \parallel_A^2 \tag{3.9}$$

donde la variable $r_{ij} \in \{0, 1\}$ indica la pertenencia de un elemento a un subconjunto; por lo tanto, r_{ij} =1 si el elemento x_i pertenece al subconjunto H_j y r_{ij} =0 en caso contrario (Bishop, 2006). Para realizar tal partición, el algoritmo utiliza una técnica de refinamiento iterativo que consta de dos etapas:

1. Inicialización: Se eligen aleatoriamente k elementos de X que conformarán las semillas iniciales de los k subconjuntos o cluster.
2. Optimización: Esta etapa a su vez puede dividirse en otras dos:
   a) La etapa donde se minimiza J respecto a r_{ij} dejando fijo los valores de H_j. Es decir, se actualizan los valores de r_{ij} de tal manera que cada muestra x_n se asigne al subconjunto con mayor similitud.
   b) La etapa de maximización donde se minimiza J respecto a H_j dejando fijo los valores de r_{ij}. En otras palabras, se recalculan los valores de H_j una vez se conocen los elementos que contiene

   y se repite hasta que no haya cambios en los subconjuntos.

3.5. Consideraciones Finales

El reconocimiento de acciones con información multimodal es una nueva tendencia en el reconocimiento de acciones por lo que la explicación de estos conceptos básicos llegan a ser muy importantes. Los descriptores explicados STIP, HOG, MFCC y Espectograma fueron utilizados para el método propuesto, así como el clasificador SVM.

Capítulo 4

Propuesta

En el presente capitulo se describe el método propuesto, el cual recibe información multimodal (intensidad, profundidad y audio) y se divide en tres partes. Primero se **extrae las características** lo cual se aplica en los tres canales de información. Luego, se procede a utilizar la técnica de **Bolsa de Palabras (*bag-of-words*)** para cada tipo de fuente de datos, generando así tres diccionarios. A partir de los cuales serán calculados los histogramas de palabras visuales y de audio. Finalmente, estos atributos entraran a la etapa de **Clasificación** , donde es usado el clasificador SVM
.

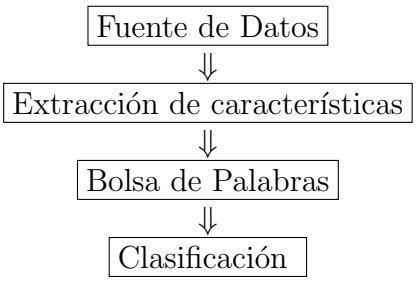


Figura 4.1: Modelo de Propuesta.

4.1. Extracción de características

4.1.0.1. Extracción de características

Debido a que se utiliza información multimodal, se realiza la extracción de características de los distintos canales (intensidad, profundidad y audio).

Información de intensidad La extracción de características en este canal consta de dos partes, la detección de puntos de interés y la descripción de los mismos Todos los cuadros (*frames*) de un video son tratados juntos, generando una forma tridimensional, la cual sirve de entrada al algoritmo STIP Laptev (2005). En primer lugar, se realiza la detección de puntos de interés, al realizar su descripción terminan siendo un vector numérico que describen en forma matemática las características.

Dicho algoritmo detecta puntos de interés espacio-temporales, identifica aquellas regiones con mayor variación de grises utilizando una variación del detector de esquinas Harris 3D. Una vez detectados los puntos de interés se procede a describir los mismos usando el histograma de orientación de gradientes (HOG) y el histograma de flujo óptico (HOF), en ambos casos los histogramas se definen por 72 dimensiones logrando asi un una matriz de **144 x n**.

Información de profundidad De igual forma la extracción de características consta de dos partes: la detección de puntos de interés y la descripción de los mismos, en este caso usamos el descriptor HOG presentado en Dalal y Triggs (2005). El presente descriptor transforma una imagen a componentes "básicos"que representen a la imagen original. Consiste en tres etapas:calcular los vectores de los gradientes, calcular los histogramas sobre las orientaciones de los gradientes, normalizar los gradientes.

En lugar de reducir toda la imagen de una sola vez, HOG lo hace de forma iterativa en bloques de 16×16 pixeles, es decir 256 pixeles serán reducidos a una cantidad menor de información. Luego, los histogramas generados en cada celda son normalizados. El objetivo de esta normalización local es tornar el descriptor en invariante a las variaciones de iluminación.De esta manera se obtiene una matriz de 9×n.

Información de audio En el canal de audio utilizamos dos descriptores, los Coeficientes Ceptrales en las Frecuencias de Mel (MFCC) y el Espectograma.

El MFCC se basa en estudios que aproximan la percepción auditiva humana bajo la escala de Mel, que consiste en una representación logarítmica de la señal del origen. Su calculo involucra la transformada de Fourier, el paso del resultado a la escala de Mel, y utilizar transformada discreta de coseno para finalmente, retornar las amplitudes obtenidas que corresponden a los MFCC.

El espectograma consiste en una serie de valores que expresan la relación que existe entre el espectro de potencia de una señal respecto a una señal de ruido blanco. Un espectro puede corresponder a un impulso de señal o ruido. Donde los valores de los coeficientes cuyo valor es alto, significa o refleja ruido o que no existe un tono en particular presente en dicha banda, un valor bajo en dichos coeficientes indican una estructura armónica de espectro o un tono dentro de esa banda.

De esta manera al aplicar los dos descriptores se obtiene una matriz de $n \times 13$ que corresponde al MFCC y un se obtiene una matriz de $n \times 442$ el que corresponde al espectograma , luego e concatena de forma simple.

4.2. Bag of Words

La técnica *Bag-of-Words* (BoW) fue originalmente usada para la extracción de información de nivel medio en documentos y palabras a partir de atributos de bajo nivel. Esta técnica consta de la extracción de características, generación del diccionário y cálculo del histograma de palabras visuales o de audio,los pasos se pueden observan en Figura 4.2.

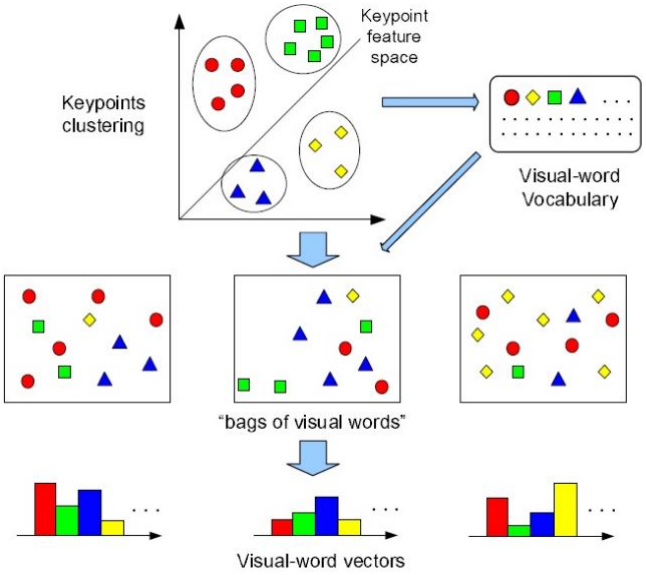


Figura 4.2: Pasos de la Bolsa de Palabras .

4.2.1. Generación de diccionario

Para la generación del diccionario se toma una muestra, en este trabajo se opto por un 10 % del total de videos pertenecientes a una base de datos. El algoritmo de agrupamiento usado en la propuesta es K-means. Con el cual se generan K grupos, cada grupo o *cluster* recibe el nombre de *codeword*, el mismo que es representado por el centroide del grupo como se observa en la Figura 4.3. Por lo tanto, cada *codeword* representa un grupo de características similares; en la presente propuesta después de varias pruebas se optó por generar diccionarios de 400 palabras.

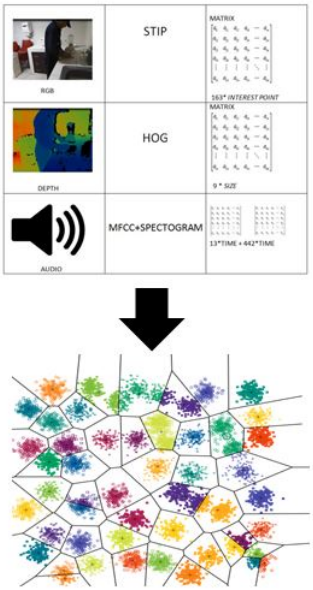


Figura 4.3: Generación de diccionarios .

4.2.2. Generación de histogramas visuales

Después de generar los diccionarios *codebook*, son creados histogramas como se observa en la Figura 4.4, los cuales cuentan las ocurrencias de cada *codeword* en la imagen o el audio, para generar estos histogramas utilizamos distancia Euclidiana.

Figura 4.4: Generación de histogramas .

4.3. Clasificación

El método usado en la clasificación es el algoritmo SVM Cortes y Vapnik (1995). Este clasificador fue seleccionado por su alta tasa de acierto,Figura 4.5. SVM consigue una buena generalización a partir de un pequeño conjunto de datos. SVM también tiene la propiedad de hacer posible la clasificación no lineal usando la teoria de kernels sin necesitar un algoritmo específico no lineal. Los kernels son usados para mapear los datos en un espacio de características de alta dimensión.

SVM ha sido establecido como un potente algoritmo de aprendizaje con una buena capacidad de generalización, lo han demostrado trabajos como (Wallraven et al., 2003) donde lograr tener mas del 90 % de reconocimiento de rostros gracias a este clasificador.

Figura 4.5: Clasificación .

4.4. Consideraciones Finales

En el presente capítulo, se ha descrito el modelo propuesto. En el mismo se propone el uso de informaciones multimodales, tales como intensidad, profundidad y audio. Será realizado un conjunto de experimentos para determinar el descriptor que mejor caracterice cada tipo de informação. La técnica de bolsa de palabras es usada en los descriptores de las tres fuentes de información. Finalmente, los histogramas generados por el técnica BoW serán las entradas a un clasificador SVM.

Capítulo 5

Pruebas y Resultados

En el presente capítulo se presentan los experimentos realizados usando la propuesta de esta tesis. Así mismo se hace una análisis de los resultados, además de una comparación con otros métodos de la literatura.

5.1. Bases de Vídeos

Consiste en un conjunto de objetos multimedia que comparten características, tales como: el tipo de contenido, el formato en el que se encuentran, etc. Las bases de vídeos son usadas para evaluar nuestra propuesta de reconocimiento de acciones humanas a través de información multimodal. Además de eso, con el uso de bases públicas podemos hacer una comparación con otras propuestas de la literatura.

5.1.1. Hollywood

Hollywood (Marszalek et al., 2009) es un conjunto de datos formados por una serie de videos ya etiquetados que consisten en partes de distintas películas que encapsulan acciones humanas. Consta de 823 videos de entrenamiento y 884 de prueba, divididos en 12 clases: *contestar el teléfono, conducir un auto, comer, pelear, bajar del auto, saludo de manos, abrazar, besar, correr, sentarse, recostarse y levantarse.* En la Figura 5.1 se observa algunas imágenes de la base de vídeos Hollywood.

5.1.2. KTH

KTH (Schuldt et al., 2004) es un conjunto de vídeos que contiene seis tipos de acciones humanas: *caminar, trotar, correr, boxeo, agitar la mano y aplaudir.* Las mismas son realizadas en varias ocasiones por 25 sujetos en cuatro escenarios diferentes:

Figura 5.1: Imágenes de la base de vídeos Hollywood.

al aire libre, al aire libre con variación de la escala, al aire libre con diferentes tipos de ropa y en el interior, como se observa e ilustra en la Figura 5.2. La base de datos contiene 2391 secuencias de vídeos. Todas las secuencias fueron tomadas sobre fondos homogéneos con una cámara estática.

Figura 5.2: Imágenes de la base de vídeos KTH.

5.1.3. CAD 120

CAD 120 (Koppula et al., 2013) es un conjunto de 120 vídeos RGB-D que consiste en actividades diarias realizadas por cuatro sujetos: dos varones y dos mujeres. Entre las mujeres, una de ellas es zurda. Las actividades realizadas en esta base son de alto nivel: *apilar objetos, desapilar objetos, tomar medicina, usar microondas, limpiar microondas, preparar cereal, organizar objetos, tomar comida de microondas, comer y buscar objeto*. Podemos observar algunas muestras de la base CAD 120 en la Figura 5.3.

Figura 5.3: Imágenes de la base de vídeos CAD120.

5.1.4. Human Manipulation Actions (HMA)

La base de vídeos HMA (Pieropan et al., 2014b) es un conjunto de vídeos donde ocho sujetos realizan la tarea de preparar cereales para el desayuno. A los actores no se les instruye sobre cómo realizar la acción. Las acciones se dividen en: *abrir leche, servir leche, cerrar leche, abrir cereal, servir cereal y cerrar cereal*. En la Figura 5.4 se presenta un ejemplo de cuadros de intensidad y de profundidad de la base HMA. En la primera fila tenemos el cuadro de intensidad y en la siguiente linea el respectivo cuadro de profundidad.

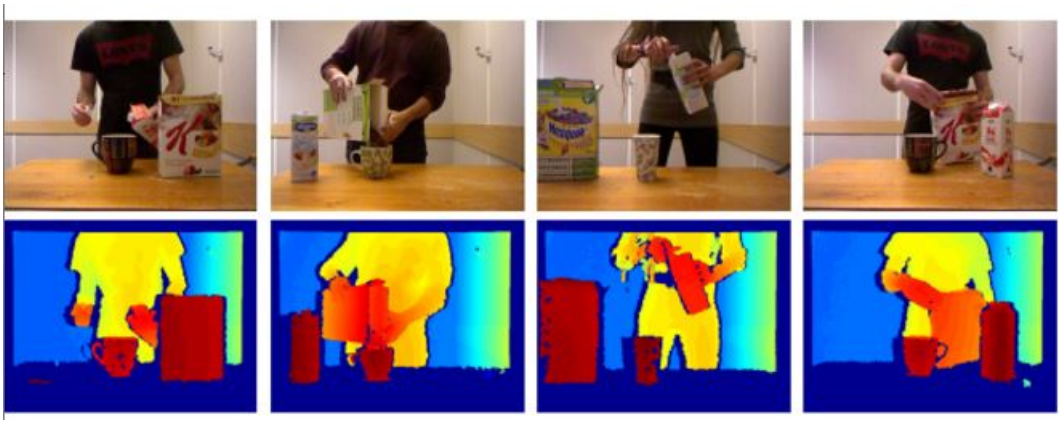


Figura 5.4: Imágenes de la base de vídeos Human Manipulation Actions

5.1.5. Kitchen-UCSP

Kitchen-UCSP es un conjunto de vídeos propios donde diez personas realizan tareas comunes que ocurren en la cocina. Acciones como: *apagar luz, usar cuchillo eléctrico, abrir frasco, golpear, lavarse las manos, usar licuadora, usar microondas, picar, rallar pan y secarse manos*. En la Figura 5.5, se muestra un ejemplo de cuadros de intensidad y profundidad de la base Kitchen-UCSP.

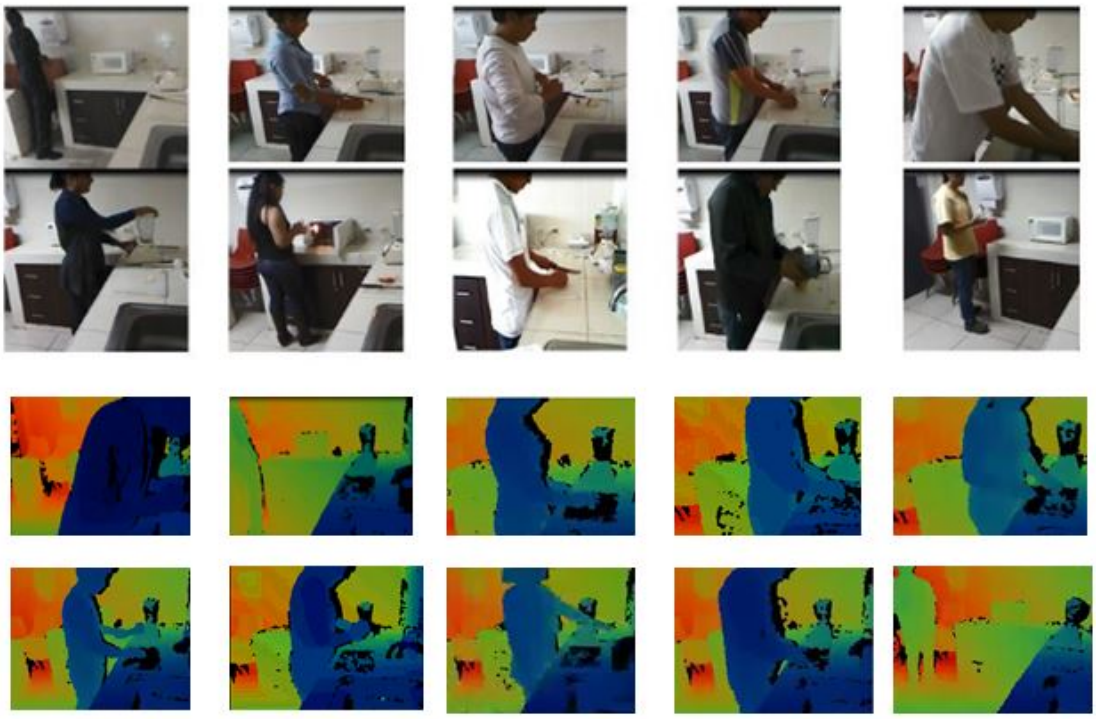


Figura 5.5: Imágenes de la base de vídeos Kitchen-UCSP

5.2. Métricas

Para la etapa de experimentación se requieren maneras de medir el desempeño del método propuesto, específicamente la eficacia, donde se mide la capacidad de retornar la clase correcta de los vídeos consultados.

5.2.1. Matriz de confusión

Una matriz de confusión es una herramienta que permite la visualización del desempeño de un algoritmo que se emplea en aprendizaje supervisado. Cada columna de la matriz representa el número de predicciones de cada clase, mientras que cada fila representa a las instancias en la clase real. Uno de los beneficios de las matrices de

confusión es que facilitan ver si el sistema está confundiendo una clase con otra, como se observa en el Cuadro 5.1.

Clases	Apilar	Desapilar
Apilar	**0.8**	0.2
Desapilar	0.4	**0.6**

Cuadro 5.1: Matriz de Confusión.

En el Cuadro 5.1 se puede observar que la clase *Apilar* obtiene un 80 % de acierto y confunde la acción con la clase *Desapilar* un 20 % de la totalidad. La clase *Desapilar* obtiene un 60 % de acierto y confunde con la clase *Apilar* un 40 %, según el cuadro se puede observar que existe dificultad en el reconocimiento de la acción *Desapilar* ya que hay una fuerte confusión con la acción *Apilar*.

5.2.2. Definición de parámetros

Porcentaje de entrenamiento y de test: Para todos los experimentos, se considero el 50 % del total de las bases de vídeos para el entrenamiento y 50 % para el test. Cada experimento se realizó en promedio 5 veces y se considera el promedio de todos los resultados existentes.

Formación de diccionarios: Para la formación de diccionarios se utilizó el algoritmo de agrupamiento K-means donde el número centroides varía según el tipo de información. Para el canal de intensidad y de profundidad se consideran 500 centroides y para el canal de audio 350, se llego a este numero, después de varias pruebas las cuales se muestran a continuación.

1. **Intensidad:** En el canal de intensidad se considero 500 centroides para la generación del diccionario, esto se debe a las diferentes pruebas que se realizaron. Se utilizó la base de vídeos CAD120. En el Cuadro 5.2 se observa los resultados de las diferentes pruebas con distintos números de centroides siendo en promedio, el mejor de los resultados el de 500 centroides.

2. **Profundidad:** En el canal de profundidad, de manera similar al de intensidad, se considero 500 centroides para la generación del diccionario. El tamaño del diccionario se determinó a través de varias pruebas con la base de vídeos CAD120. En el Cuadro 5.3 se muestra los resultados que se realizaron con diferentes números de centroides, siendo en promedio el mejor de los resultados el de 500 centroides.

3. **Audio:** En el canal de audio se considero 350 centroides para la generación del diccionario, esto se debe a varias pruebas que se realizaron en la base de vídeos

Acciones	Número de clusters				
	200	**300**	**400**	**500**	**700**
Apilar	0.33	0.33	0.50	**0.50**	0.50
Desapilar	0.0	0.66	0.50	**0.66**	0.62
Tomando medicina	0.33	0.83	0.83	**1.00**	0.83
Usar microondas	0.75	1.00	0.83	**0.81**	0.75
Limpiar microondas	0.16	0.00	0.50	**0.64**	0.50
Preparar cereal	0.60	0.20	0.50	**0.64**	0.62
Organizar objetos	0.16	0.50	0.61	**0.61**	0.50
Tomar comida de microondas	0.12	0.12	0.62	**0.65**	0.50
Comiendo	0.50	0.25	0.75	**1.00**	0.75
Buscar objeto	0.83	1.00	0.66	**0.82**	0.66
Promedio general	0.38	0.49	0.69	**0.73**	0.62

Cuadro 5.2: Tabla comparativa sobre la asertividad del canal de intensidad en la base de vídeos CAD120.

Acciones	Número de clusters				
	200	**300**	**400**	**500**	**700**
Apilar	0.33	0.50	0.33	**0.50**	0.33
Desapilar	0.0	0.33	0.50	**0.66**	0.66
Tomando medicina	0.33	0.50	0.66	**0.66**	0.50
Usar microondas	0.75	0.50	0.50	**0.75**	0.50
Limpiar microondas	0.16	0.33	0.33	**0.66**	0.50
Preparar cereal	0.60	0.50	0.60	**0.50**	0.60
Organizar objetos	0.16	0.33	0.50	**0.66**	0.50
Tomar comida de microondas	0.12	0.33	0.25	**0.62**	0.25
Comiendo	0.50	0.50	0.50	**0.50**	0.50
Buscar objeto	0.83	0.50	0.83	**0.83**	0.83
Promedio general	0.38	0.43	0.50	**0.63**	0.52

Cuadro 5.3: Tabla comparativa sobre la asertividad del canal de profundidad en la base de vídeos CAD120.

HMA el Cuadro 5.4 hace referencia a los resultados, se realizó la prueba con diferentes números de centroides siendo en promedio el mejor de los resultados el de 350 centroides.

5.3. Resultados

Se realizaron varios experimentos, usando el modelo de esta propuesta con diversas bases de vídeos con uno, dos y tres canales de información.

Acciones	**Número de clusters**					
	100	**200**	**300**	**350**	**400**	**500**
abrir cereal	0.53	0.47	0.39	**0.78**	0.73	0.47
servir cereal	0.11	0.77	0.73	**1.00**	0.89	0.93
cerrar cereal	0.20	0.25	0.66	**0.80**	0.52	0.34
abrir leche	0.51	0.45	0.40	**0.85**	0.51	0.70
servir leche	0.5	0.28	0.77	**1.00**	0.88	0.84
cerrar leche	0.0	0.34	0.47	**0.36**	0.70	0.43
Promedio general	0.23	0.43	0.57	**0.80**	0.71	0.62

Cuadro 5.4: Tabla comparativa sobre la asertividad del canal de audio en la base de vídeos HMA.

5.3.1. Experimento 1: Intensidad

En el canal de intensidad se hicieron pruebas con las siguientes bases de vídeos:

Hollywood: Se aplicó el método propuesto sobre esta base. El descriptor utilizado fue STIP, se consideró solo cinco acciones: *contestar teléfono, dar la mano, abrazar, pararse y levantarse.* Las acciones *abrazar* y *pararse* logran una asertividad del 80 %. Los resultados se muestran en el Cuadro 5.5. Con base en el experimento, se puede afirmar que STIP tiene buen desempeño en imágenes de intensidad. Se obtuvo un porcentaje de asertividad del 74.06 % a comparación con métodos similares, es el más alto como se observa en el Cuadro 5.6.

	Contestar telefono	Dar la mano	Abrazar	Pararse	Levantarse
Contestar telefono	**0.76**	0.12	0.12	0.0	0.0
Dar la mano	0.20	**0.68**	0.12	0.0	0.0
Abrazar	0.05	0.15	**0.80**	0.0	0.0
Pararse	0.0	0.0	0.0	**0.80**	0.20
Levantarse	0.0	0.05	0	0.28	**0.67**

Cuadro 5.5: Matriz de confusión usando información de intensidad en la base Hollywood

El incremento de asertividad en el Cuadro 5.6, se debe a que los vídeos son divididos en clips, también se debe notar que acciones como *contestar teléfono, dar la mano y abrazar* se llegan a confundir entre sí, esto se debe a la similitud de la posición del cuerpo al ejecutar la acción. El porcentaje de asertividad en esta base de vídeos es baja comparada con los resultados de otras bases de vídeos, esto se debe a que contiene en su mayoría fondos dinámicos y muchas veces la acción ocupa un espacio muy pequeño comparado con el tamaño del *frame*.

Acciones	**Investigaciones**		
	STIP (Laptev et al., 2008b)	SIFT (Kulkarni et al., 2015)	**Propuesta**
Contestar teléfono	0.32	0.15	0.75
Dar la mano	0.32	0.20	0.67
Abrazar	0.40	0.37	0.80
Pararse	0.18	0.23	0.80
Levantarse	0.5	0.53	0.66

Cuadro 5.6: Comparación de tasas de acierto usando información de intensidad en la base Hollywood.

KTH: Se aplicó el método propuesto sobre la base de videos KTH, los vídeos de esta base solo poseen información de intensidad de color. El descriptor utilizado con esta base fue STIP, obteniendo más del 95 % en cinco de las seis clases de esta base de vídeos. Los resultados se muestran en el Cuadro 5.7. Basándonos en el experimento, podemos afirmar que STIP tiene buen desempeño en imágenes de intensidad ya que en los experimentos realizados en la investigación, como en otras investigaciones han llegado a resultados similares como se indica en el Cuadro 5.8.

	Caminar	Trotar	Correr	Boxeo	Agitar mano	Aplaudir
Caminar	**0.98**	0.02	0.0	0.0	0.0	0.0
Trotar	0.12	**0.67**	0.18	0.0	0.0	0.03
Correr	0.0	0.0	**1.00**	0.0	0.0	0.0
Boxeo	0.0	0.0	0.0	**1.00**	0.0	0.0
Agitar mano	0.0	0.0	0.0	0.0	**0.87**	0.13
Aplaudir	0.0	0.0	0.0	0.0	0.05	**0.95**

Cuadro 5.7: Matriz de confusión usando información de intensidad en la base KTH.

	Investigaciones			
	Veeriah et al. (2015)	Wang et al. (2010)	Gilbert et al. (2011)	**Propuesta**
Asertividad	0.93 %	0.95 %	0.91 %	0.95 %

Cuadro 5.8: Comparación de tasas de acierto usando información de intensidad en la base KTH.

El promedio de asertividad en esta base de vídeos es más alta que en el experimento anterior, esto se debe a que la base de vídeos posee fondos homogéneos y las actividades son muy distintas respecto a la posición del cuerpo, todas las actividades se realizan en dos escenarios: al aire libre y en un ambiente cerrado, esto facilita el reconocimiento de la acción.

5.3.2. Experimento 2: Intensidad y Profundidad

CAD120 Se aplicó el método propuesto sobre la base de videos Cad 120, la que contiene información de dos canales (intensidad y profundidad). Los descriptores utilizados con esta base fueron STIP y HOG, ambos descriptores fueron aplicados en los dos canales de información. El Cuadro 5.9 muestra los resultados obtenidos, fueron evaluados los canales de información de intensidad y de profundidad. Se obtuvo un acierto del 100 % de asertividad en acciones como: *tomar medicinas, comer y preparar cereal*. Ya en la acción *limpiar microonda*, podemos observar que el descriptor HOG tuvo un impacto importante en la mejora de los resultados. En la acción *comiendo* el descriptor STIP alcanzó uma mayor taza de acierto, esto ocurre debido al movimiento repetitivo y estructurado que existe en esa acción. Lo que comprometió el resultado del descriptor STIP en los datos de profundidad es la cantidad de ruido que es generado en la captura de la información. Basados en los resultados, podemos afirmar que el descriptor STIP tiene un mejor desempeño en imágenes de intensidad y el descriptor HOG en imágenes de profundidad, como lo muestra la Figura 5.6.

Acción	**RGB-D (STIP)**	**RGB-D (HOG)**	**RGB (STIP) Depth (HOG)**	**RGB (HOG) Depth (STIP)**
Apilar	0.50	0.35	**0.55**	0.41
Desapilar	0.20	0.81	**0.85**	0.43
Tomando medicina	1.00	1.00	**1.00**	1.00
Usar microondas	0.64	0.81	**0.85**	0.68
Limpiar microondas	0.64	1.00	**1.00**	0.78
Preparar cereal	0.61	1.00	**1.00**	0.78
Organizar objetos	0.65	0.81	**0.85**	0.70
Tomar comida de microondas	0.60	0.35	**0.73**	0.44
Comiendo	**1.00**	0.82	0.94	0.90
Buscar objeto	0.82	0.82	**0.90**	0.80

Cuadro 5.9: Comparación de diferentes canales en la base de vídeos CAD-120.

En el Cuadro 5.10 se observa la comparación con los creadores de la base de vídeos, donde utilizan diferentes tipos de etiquetas para la clasificación, etiquetado simple, cuando no hay divisiones del vídeo, *i.e.*, cuando existe una acción atómica en el video; y etiquetado de alto nivel cuando dividen el vídeo en sub-acciones.

	Investigaciones		
	(Koppula et al., 2013)	(Koppula et al., 2013)	**Propuesta**
Asertividad	0.76 %	0.84 %	0.86 %

Cuadro 5.10: Comparación de tasas de acierto usando información multimodal de intensidad y profundidad en la base CAD120.

El porcentaje de asertividad en esta base de vídeos en promedio no es mucho más alta que la encontrada en la literatura actual. El método propuesto divide los vídeos en

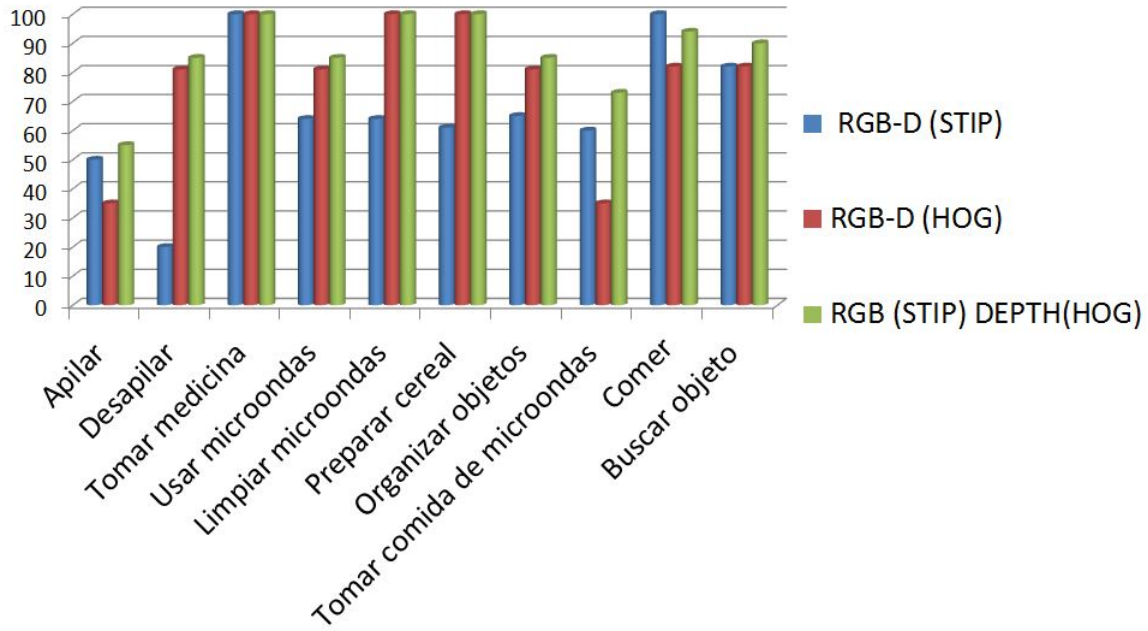


Figura 5.6: Tabla de comparación de los descriptores STIP y HOG aplicados en dos fuentes de información: intensidad y profundidad en la base de vídeos CAD-120 .

clips, los creadores de la base de vídeos dividen el vídeo en la etapa de entrenamiento dando varias etiquetas a un vídeo, lo cual llaman etiquetado de alto nivel.

5.3.3. Experimento 3 : Intensidad, Profundidad y Audio

HMA La base de vídeos posee tres fuentes de información (intensidad, profundidad y audio), necesitamos de estas tres fuentes de información para poder evaluar nuestra propuesta. Los descriptores utilizados son: STIP en el canal de intensidad, HOG en el canal de profundidad y finalmente MFCC y espectrograma en el canal de audio.

Primero se evaluó la información proveniente del canal de intensidad. El Cuadro 5.11 muestra la matriz de confusión usando solamente la información de intensidad, dicho cuadro muestra el porcentaje de acierto por clase. El promedio general de la tasa de acierto fue del 57 %, se observa que las clases *servir leche* y *cerrar leche* son las que tienen peor desempeño ya que visualmente son muy semejantes, incluso para un ser humano resulta difícil diferenciar algunas acciones y se confunden entre si.

Luego se evaluó el canal de profundidad. El Cuadro 5.12 muestra la matriz de confusión para este canal. Comparando el promedio general de la tasa de acierto de esta fuente con el canal de intensidad, hubo un incremento de acierto del 3 %, logrando de esta forma una tasa del 60 %. Aún persiste los bajos resultados con las clases *servir leche* y *cerrar leche*. Haciendo una comparación con los resultados de la matriz de confusión de intensidad, podemos ver que hubo un incremento en la tasa de acierto de las clases *abrir, cereal, servir cereal* y *cerrar cereal*.

	Abrir cereal	Servir cereal	Cerrar cereal	Abrir leche	Servir leche	Cerrar leche
Abrir cereal	**0.68**	0.00	0.09	0.23	0.00	0.00
Servir cereal	0.00	**0.80**	0.00	0.00	0.20	0.00
Cerrar cereal	0.15	0.00	**0.68**	0.00	0.00	0.17
Abrir leche	0.00	0.00	0.00	**0.85**	0.00	0.15
Servir leche	0.08	0.05	0.25	**0.33**	0.18	0.11
Cerrar leche	0.07	0.10	**0.30**	**0.30**	0.10	0.13

Cuadro 5.11: Matriz de confusión usando información de intensidad en la base HMA.

	Abrir cereal	Servir cereal	Cerrar cereal	Abrir leche	Servir leche	Cerrar leche
Abrir cereal	**0.78**	0.00	0.10	0.12	0.00	0.00
Servir cereal	0.00	**0.89**	0.00	0.00	0.11	0.00
Cerrar cereal	0.20	0.00	**0.80**	0.00	0.00	0.00
Abrir leche	0.00	0.00	0.00	**0.85**	0.00	0.15
Servir leche	0.00	0.00	0.13	0.33	0.18	**0.36**
Cerrar leche	0.00	0.00	0.16	**0.50**	0.21	0.13

Cuadro 5.12: Matriz de confusión usando información de profundidad en la base HMA.

También se evaluó el canal de audio. El Cuadro 5.13 muestra los resultados usando este canal como fuente información. Gracias al audio fue posible reconocer acciones que con fuentes visuales no era posible reconocer. Por ejemplo, era difícil reconocer la clase *servir leche* usando solo información visual. A través del audio se logró identificar el sonido que se emite cuando la leche cae sobre un recipiente, permitiendo que esta acción sea fácilmente identificada. La tasa promedio de acierto con el audio fue del 80 %, logrando 100 % de acierto en las clases *servir cereal* y *servir leche*. Como se puede observar en los resultados el aporte del sonido fue significativo para el reconocimiento de algunos tipos de acciones. La clase *cerrar leche* aún tiene un tasa de acierto baja, eso ocurre porque el sonido de destapar y tapar es el mismo, solo que uno es en un sentido y el otro es en el contrario.

	Abrir cereal	Servir cereal	Cerrar cereal	Abrir leche	Servir leche	Cerrar leche
Abrir cereal	**0.78**	0.00	0.10	0.12	0.00	0.00
Servir cereal	0.00	**1.00**	0.00	0.00	0.00	0.00
Cerrar cereal	0.20	0.00	**0.80**	0.00	0.00	0.00
Abrir leche	0.00	0.00	0.00	**0.85**	0.00	0.15
Servir leche	0.00	0.00	0.00	0.00	**1.00**	0.00
Cerrar leche	0.00	0.00	0.00	**0.64**	0.00	0.36

Cuadro 5.13: Matriz de confusión usando información de audio en la base HMA.

Finalmente, cuando son utilizados los 3 canales de información se obtiene un promedio de 88 % de acierto, siendo el de menor porcentaje el de *cerrar leche* ya que

es difícil diferenciarlo con la clase *abrir leche* debido a que son muy semejantes, incluso un ser humano tendría la misma dificultad, logrando un 66 %, siendo este más alto que el promedio general de solo el canal de intensidad y profundidad. Los resultados son mostrados en el Cuadro 5.14. En todas las clases hubo una mejora en las tasas de acierto. Entonces, podemos concluir que los tres tipos de información se complementan entre si, permitiendo conseguir tasas de acierto mayores. Por lo tanto, no podemos descartar las informaciónes provenientes de otras fuentas, ya que con el uso de todas ellas se podrá mejor la tasa de acierto.

	Abrir cereal	Servir cereal	Cerrar cereal	Abrir leche	Servir leche	Cerrar leche
Abrir cereal	**0.90**	0.00	0.10	0.00	0.00	0.00
Servir cereal	0.00	**0.98**	0.00	0.02	0.00	0.00
Cerrar cereal	0.09	0.00	**0.91**	0.00	0.00	0.00
Abrir leche	0.00	0.00	0.00	**0.85**	0.00	0.15
Servir leche	0.00	0.00	0.02	0.00	**0.98**	0.00
Cerrar leche	0.00	0.00	0.00	0.34	0.00	**0.66**

Cuadro 5.14: Matriz de confusión usando los tres canales de información (intensidad, profundidad y audio) en la base HMA.

A diferencia del método original planteado por Pieropan et al. (2014a), el cual considera al silencio como una acción mas, la forma como se aplica la propuesta, es segmentado el vídeo en clips y procesando cada uno de ellos en forma independiente. En la Figura 5.15 se compara el método propuesto con los resultados de los propios creadores de la base de vídeos. Cabe indicar que el método de la literatura también hizo uso de los tres canales de información.

	Pieropan	Método Pro-puesto
Abrir cereal	**0.90**	0.60
Servir cereal	**0.98**	0.82
Cerrar cereal	**0.91**	0.37
Abrir leche	**0.83**	0.42
Servir leche	**0.98**	0.64
Cerrar leche	**0.62**	0.38

Cuadro 5.15: Comparación del metodo propuesto con los creadores de la base de videos..

En el cuadro 5.16 observamos la comparación de la asertividad promedio con los creadores de la base de vídeos, debemos recalcar que para obtener estos resultados se realiza una pre-segmentación en los vídeos, quedando divididos en clips, los creadores de la base realizan una segmentación continua donde conectan el estado final de cada sub-acción con el estado inicial de la siguiente. También consideran como una sub-acción el ruido a la cual la etiquetan como *Garbage*. El reconocimiento que utilizan se lleva a cabo mediante la búsqueda del camino más probable a través del algoritmo

de Viterbi, en la fusión utilizan al igual que nosotros una de bajo nivel mediante la definición que es la concatenación de características de audio y vídeo.

En la literatura actual diversas técnicas de reconocimiento de acciones se han centrado en diferentes modalidades individuales de las señales. Para un mejor rendimiento del reconocimiento, es deseable fusionar esta información multimodal en un conjunto integrado de características discriminantes.

Un problema interesante que se plantea consiste en cómo una entrada multi-modal puede ser integrada para formular una interpretación coherente de una escena. Lo ideal sería que dicha fusión debería mitigar las debilidades de las fuentes individuales. La fusión puede realizarse a cualquier nivel en el proceso de aprendizaje, cada método tiene sus puntos fuertes y débiles, en este trabajo se utilizó el nivel bajo de fusión mediante la concatenación de los vectores de características.

Se extrajeron los coeficientes (MFCC). Este es uno de los conjuntos más sólidos y ampliamente utilizados en el campo de la extracción de características basadas en audio. MFCC fue diseñado principalmente para el reconocimiento de voz, pero hay un gran número de trabajos en los que se han utilizado para clasificar a un amplio conjunto de diferentes clases de sonido, por este motivo también el uso del espectrograma para este canal y así el vector resultante sea mucho mas robusto.

Esto junto a la pre-segmentación logran una metodología robusta,logrando así un porcentaje de asertividad alto, aunque en acciones muy similares como *abrir leche , cerrar leche* o acciones como *abrir cereal, cerrar cereal* aún se tiene un gran porcentaje de confusión y se debe a la similitud de acciones, ya que estas son difíciles de diferenciar hasta para el ser humano.

	Investigaciones	
	(Pieropan et al., 2014b)	**Propuesta**
Asertividad	0.73 %	0.88 %

Cuadro 5.16: Comparación de tasas de acierto usando información multimodal intensidad y profundidad en la base HMA.

Kitchen-UCSP Usando la base de vídeos Kitchen-UCSP, se evaluó la información por cada canal y con todos los canales usando así la información multimodal. En el Cuadro 5.17 se muestra las clases de esta base de vídeos y sus respectivas abreviaturas para un mejor entendimiento.

EL Cuadro 5.18 muestra los resultados usando el canal de audio como fuente información, llegando a tener como máximo un 94 % en acciones como *usar licuadora y usar microondas*. La tasa promedio de acierto con el audio fue del 80 % de acierto, clases como *usar licuadora y usar cuchillo eléctrico* se confunden entre si, debido al sonido que producen estas acciones ya que son muy similares.

Clase	Abreviatura
Apagar luz	AL
Usar cuchillos eléctrico	CE
Abrir frasco	FR
Golpear	GO
Lavarse las manos	LM
Usar Licuadora	LI
Usar Microondas	MI
Picar	PI
Rallar pan	RP
Secarse las manos	SM

Cuadro 5.17: Descripción y abreviaturas de la base de vídeos Kitchen-UCSP.

De la misma manera se evaluó el canal de intensidad. El Cuadro 5.19 muestra los resultados, usando este canal se llego a tener como máximo un 83.3 % en acciones como *usar licuadora, usar microondas y secarse las manos*. La tasa promedio de acierto con la intensidad fue del 73.3 % de acierto, clases como *golpear y picar* se confunden entre sí debido a que la postura que adopta el cuerpo es muy similar en ambas acciones.

También se evaluó el canal de profundidad. El Cuadro 5.20 muestra los resultados usando este canal, se llegó a tener como máximo un 83,3 % en la acción como *secarse las manos*. La tasa promedio de acierto con profundidad es de 75 % de acierto, superando los resultados de solo intensidad. La clase *Apagar luz* subió un 15 % debido a que este canal es invariante a la luz.

Por último se evaluó los tres canales juntos, superando así cualquier resultado anterior. El Cuadro 5.21 muestra los resultados usando estos canales, se llegó a tener como máximo un 94,4 % en la acción *secarse manos*. La tasa promedio de acierto con los tres canales es de 86.11 %.

Se debe tomar en cuenta que hay acciones, donde el cuerpo toma la misma posición, como las clases *usar cuchillo eléctrico, picar, golpear* ya que la diferencia de estas acciones está en el movimiento del brazo lo que llega a confundir al clasificador, pasa lo mismo con sonidos similares como el de las clases *usar licuadora y usar cuchillo eléctrico*.

5.4. Consideraciones Finales

En el presente capítulo se aplicó el método propuesto en diferentes bases de vídeos, en las cuales se obtuvo buenos porcentaje de asertividad ya que el uso de información multimodal *(intensidad profundidad y audio)* hace mucho mas robusto el método; Se debe considerar que la mayoría de acciones que se confunden entre si, son muy similares , generando prácticamente el mismo tipo de movimiento, con una ligera variación,

además el canal de audio es muy similar lo que hace muy difícil el reconocimiento hasta para el ser humano, un punto importante es la formación del diccionario el cual a más número de centroides más largo el proceso. El método propuesto utiliza distintos descriptores para diferentes tipos de información logrando que el rendimiento de este sobre las diferentes bases de vídeos muestran que el método propuesto obtiene buenos resultados.

	AL	CE	FR	GO	LM	LI	MI	PI	RP	SM
AL	**0.67**	0.00	0.22	0.11	0.00	0.00	0.00	0.00	0.00	0.00
CE	0.00	**0.83**	0.00	0.00	0.00	0.11	0.06	0.00	0.00	0.00
FR	0.00	0.00	**0.77**	0.00	0.00	0.00	0.00	0.06	0.11	0.06
GO	0.00	0.00	0.00	**0.83**	0.00	0.00	0.00	0.17	0.00	0.00
LM	0.00	0.28	0.00	0.00	**0.66**	0.00	0.06	0.00	0.00	0.00
LI	0.00	0.06	0.00	0.00	0.00	**0.94**	0.00	0.00	0.00	0.00
MI	0.00	0.06	0.00	0.00	0.00	0.00	**0.94**	0.00	0.00	0.00
PI	0.00	0.00	0.28	0.00	0.00	0.00	0.00	**0.61**	0.00	0.11
RP	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**0.83**	0.17
SM	0.00	0.00	0.00	0.11	0.00	0.00	0.00	0.00	0.00	**0.89**

Cuadro 5.18: Matriz de confusión usando información de audio en Base de vídeos Kitchen-UCSP.

	AL	CE	FR	GO	LM	LI	MI	PI	RP	SM
AL	**0.61**	0.00	0.11	0.00	0.00	0.00	0.17	0.11	0.00	0.00
CE	0.00	**0.72**	0.11	0.11	0.00	0.00	0.00	0.06	0.00	0.00
FR	0.00	0.00	**0.66**	0.11	0.00	0.00	0.00	0.06	0.11	0.06
GO	0.00	0.00	0.00	**0.66**	0.00	0.06	0.00	0.22	0.02	0.00
LM	0.00	0.06	0.00	0.00	**0.72**	0.00	0.06	0.00	0.16	0.00
LI	0.00	0.06	0.00	0.00	0.00	**0.83**	0.00	0.06	0.05	0.00
MI	0.00	0.06	0.00	0.00	0.00	0.00	**0.83**	0.00	0.00	0.11
PI	0.00	0.11	0.00	0.06	0.00	0.00	0.00	**0.72**	0.00	0.11
RP	0.00	0.00	0.11	0.06	0.00	0.00	0.00	0.11	**0.72**	0.00
SM	0.00	0.00	0.00	0.00	0.00	0.00	0.17	0.00	0.00	**0.83**

Cuadro 5.19: Matriz de confusión usando información de intensidad en Kitchen-UCSP.

	AL	CE	FR	GO	LM	LI	MI	PI	RP	SM
AL	**0.77**	0.00	0.00	0.00	0.00	0.00	0.17	0.06	0.00	0.00
CE	0.00	**0.66**	0.00	0.22	0.00	0.00	0.00	0.22	0.00	0.00
FR	0.00	0.00	**0.77**	0.12	0.00	0.00	0.00	0.11	0.00	0.00
GO	0.00	0.00	0.00	**0.72**	0.00	0.06	0.00	0.16	0.06	0.00
LM	0.00	0.06	0.00	0.00	**0.72**	0.00	0.06	0.00	0.16	0.00
LI	0.00	0.06	0.00	0.00	0.00	**0.77**	0.00	0.11	0.06	0.00
MI	0.00	0.06	0.00	0.00	0.06	0.00	**0.77**	0.00	0.00	0.11
PI	0.00	0.06	0.00	0.06	0.00	0.00	0.00	**0.77**	0.00	0.11
RP	0.00	0.00	0.06	0.06	0.00	0.00	0.00	0.11	**0.77**	0.00
SM	0.00	0.00	0.00	0.00	0.00	0.00	0.17	0.00	0.00	**0.83**

Cuadro 5.20: Matriz de confusión usando información de profundidad en Base de vídeos Kitchen-UCSP.

	AL	CE	FR	GO	LM	LI	MI	PI	RP	SM
AL	**0.94**	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.00	0.00
CE	0.00	**0.83**	0.00	0.00	0.00	0.00	0.11	0.06	0.00	0.00
FR	0.00	0.00	**0.88**	0.00	0.00	0.00	0.00	0.12	0.00	0.00
GO	0.00	0.00	0.00	**0.83**	0.00	0.00	0.00	0.11	0.06	0.00
LM	0.00	0.00	0.00	0.00	**0.83**	0.00	0.00	0.06	0.00	0.11
LI	0.00	0.06	0.00	0.00	0.00	**0.77**	0.00	0.11	0.06	0.00
MI	0.00	0.06	0.00	0.00	0.06	0.00	**0.94**	0.00	0.00	0.00
PI	0.00	0.11	0.00	0.06	0.00	0.00	0.00	**0.83**	0.00	0.00
RP	0.00	0.00	0.06	0.06	0.00	0.00	0.00	0.05	**0.83**	0.00
SM	0.00	0.00	0.00	0.00	0.00	0.00	0.12	0.00	0.00	**0.88**

Cuadro 5.21: Matriz de confusión usando información de los tres canales en Base de vídeos Kitchen-UCSP.

Capítulo 6

Conclusiones y Trabajos Futuros

El reconocimiento de acciones humanas en vídeos es un tema activo que en los últimos años ha adquirido mayor interés por parte de la comunidad científica. Debido a que el vídeo es una fuente multimodal, es necesario desarrollar descriptores que permitan caracterizar todas las fuentes de información que son proveídas. Este trabajo presenta un modelo de reconocimiento de acciones haciendo uso de información multimodal (intensidad, profundidad, audio)

- Se obtuvo en promedio una asertividad del 88 % en la base de vídeos HMA cuando se considera tres canales de información. Nuestra propuesta trabaja sobre clips previamente segmentados y consigue identificar a cual tipo de clase pertenece cada uno de los clips.
- Descriptores de características locales para el reconocimiento de acciones han llegado a ser populares, y enfoques como *Bag-of-Words* han demostrado ser un modelo eficaz para el reconocimiento de acciones. Esto es debido a su capacidad para hacer frente a las variaciones en el tiempo y espacio.
- Se creó la base de vídeos Kitchen-UCSP, el consta de 10 diferentes acciones que son realizadas en una cocina. Los experimento realizados con esta base consiguen alcanzar una tasa de acierto de 86 % usando tres fuentes de información: intensidad, profundidad y audio.
- En la etapa de experimentos el uso de los descriptores fue de mucha importancia, ya que no todos son los más adecuados para los distintos canales de información, afirmando que para el método propuesto el descriptor STIP es mucho mejor para el canal de intensidad y HOG para el de profundidad.

6.1. Trabajos futuros

A continuación se lista las posibles extensiones del modelo propuesto.

- Es importante enriquecer las características de atributos de bajo nivel extraídos de los vídeos, con un pre procesamiento de la información en los diferentes canales.El preprocesamiento de la información en los canales visuales como por ejemplo el relleno de agujeros en el canal de profundidad, brindará mucho mas calidad a esta, lo que causaría mejor tasa de asertividad, ya que en el método propuesto se llega a perder información de los mapas de profundidad.
- La implementación o el uso de otro algoritmo , para acelerar el proceso de la formación de diccionarios, mejorará mucho el costo del método propuesto .

Bibliografía

Alcântara, M. F., Moreira, T. P., et al. (2014). Real-time action recognition based on cumulative motion shapes. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2917–2921. IEEE.

Bilinski, P. y Bremond, F. (2011). Evaluation of local descriptors for action recognition in videos. In *International Conference on Computer Vision Systems*, pages 61–70. Springer.

Bishop, C. M. (2006). Pattern recognition. *Machine Learning*.

Breebaart, J. y McKinney, M. F. (2004). Features for audio classification. In *Algorithms in Ambient Intelligence*, pages 113–129. Springer.

Broggi, A., Bertozzi, M., et al. (2000). Shape-based pedestrian detection. In *Proceedings of the IEEE Intelligent Vehicles Symposium*, pages 215–220. Citeseer.

Brun, L., Percannella, G., et al. (2014). Hack: A system for the recognition of human actions by kernels of visual strings. In *Advanced Video and Signal Based Surveillance (AVSS), 2014 11th IEEE International Conference on*, pages 142–147. IEEE.

Chen, J., Kam, A. H., et al. (2005). Bathroom activity monitoring based on sound. In *International Conference on Pervasive Computing*, pages 47–61. Springer.

Chen, L., Wei, H., et al. (2011). Recognition of everyday domestic activities using a depth sensor. In *BMVC 2011 Student, Workshop*, pages 27–37.

Cortes, C. y Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.

Dalal, N. y Triggs, B. (2005). Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE.

Dollár, P., Rabaud, V., et al. (2005). Behavior recognition via sparse spatio-temporal features. In *Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on*, pages 65–72. IEEE.

Eronen, A. J., Peltonen, V. T., et al. (2006). Audio-based context recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(1):321–329.

García, C. y Gómez, I. (2012). Algoritmos de aprendizaje: Knn & kmeans.

Gilbert, A., Illingworth, J., et al. (2008). Scale invariant action recognition using compound features mined from dense spatio-temporal corners. In *Computer Vision–ECCV 2008*, pages 222–233. Springer.

Gilbert, A., Illingworth, J., et al. (2011). Action recognition using mined hierarchical compound features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(5):883–897.

Harma, A., McKinney, M. F., et al. (2005). Automatic surveillance of the acoustic activity in our living environment. In *2005 IEEE International Conference on Multimedia and Expo*, pages 4–pp. IEEE.

Harris, C. y Stephens, M. (1988). A combined corner and edge detector. In *Alvey vision conference*, volume 15, page 50. Citeseer.

Hasan, M. R., Jamil, M., et al. (2004). Speaker identification using mel frequency cepstral coefficients. *variations*, 1:4.

Herath, S., Harandi, M., et al. (2016). Going deeper into action recognition: A survey. *arXiv preprint arXiv:1605.04988*.

Hu, X., Kong, B., et al. (2007). Image recognition based on wavelet invariant moments and wavelet neural networks. In *2007 International Conference on Information Acquisition*, pages 275–279. IEEE.

Huang, Z. y Leng, J. (2010). Analysis of hu's moment invariants on image scaling and rotation. In *Computer Engineering and Technology (ICCET), 2010 2nd International Conference on*, volume 7, pages V7–476. IEEE.

Jansen, B., Temmermans, F., et al. (2007). 3d human pose recognition for home monitoring of elderly. In *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 4049–4051. IEEE.

Junejo, I. N., Dexter, E., et al. (2011). View-independent action recognition from temporal self-similarities. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(1):172–185.

Klaser, A., Marszałek, M., et al. (2008). A spatio-temporal descriptor based on 3d-gradients. In *BMVC 2008-19th British Machine Vision Conference*, pages 275–1. British Machine Vision Association.

Koppula, H. S., Gupta, R., et al. (2013). Learning human activities and object affordances from rgb-d videos. *The International Journal of Robotics Research*, 32(8):951–970.

Kulkarni, K., Evangelidis, G., et al. (2015). Continuous action recognition based on sequence alignment. *International Journal of Computer Vision*, 112(1):90–114.

Laptev, I. (2005). On space-time interest points. *International Journal of Computer Vision*, 64(2-3):107–123.

Laptev, I., Marszałek, M., et al. (2008a). Learning realistic human actions from movies. In *IEEE Conference on Computer Vision & Pattern Recognition*.

Laptev, I., Marszalek, M., et al. (2008b). Learning realistic human actions from movies. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE.

Li, W., Zhang, Z., et al. (2010). Action recognition based on a bag of 3d points. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pages 9–14. IEEE.

López, D. R., Neto, A. F., et al. (2014). Reconocimiento en-línea de acciones humanas basado en patrones de rwe aplicado en ventanas dinámicas de momentos invariantes. *Revista Iberoamericana de Automática e Informática Industrial RIAI*, 11(2):202–211.

Marszalek, M., Laptev, I., et al. (2009). Actions in context. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 2929–2936. IEEE.

Messing, R., Pal, C., et al. (2009). Activity recognition using the velocity histories of tracked keypoints. In *2009 IEEE 12th international conference on computer vision*, pages 104–111. IEEE.

Niemann, H. (2013). *Klassifikation von mustern*. springer-Verlag.

Ojala, T., Pietikainen, M., et al. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on pattern analysis and machine intelligence*, 24(7):971–987.

Peltonen, V., Tuomi, J., et al. (2002). Computational auditory scene recognition. In *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, volume 2, pages II–1941. IEEE.

Pieropan, A., Salvi, G., et al. (2014a). Audio-visual classification and detection of human manipulation actions. In *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3045–3052. IEEE.

Pieropan, A., Salvi, G., et al. (2014b). A dataset of human manipulation actions. In *IEEE International Conference on Robotics and Automation: International Workshop on Autonomous Grasping and Manipulation-An Open Challenge, Hong Kong, China, 2014*.

Poppe, R. (2010). A survey on vision-based human action recognition. *Image and vision computing*, 28(6):976–990.

Rohr, K. (1994). Towards model-based recognition of human movements in image sequences. *CVGIP: Image understanding*, 59(1):94–115.

Schuldt, C., Laptev, I., et al. (2004). Recognizing human actions: a local svm approach. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 3, pages 32–36. IEEE.

Sempena, S., Maulidevi, N. U., et al. (2011). Human action recognition using dynamic time warping. In *Electrical Engineering and Informatics (ICEEI), 2011 International Conference on*, pages 1–5. IEEE.

SENA (2009). Aspectos Básicos del Vídeo Digital.

Sheikh, Y., Sheikh, M., et al. (2005). Exploring the space of a human action. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 1, pages 144–149. IEEE.

Shotton, J., Girshick, R., et al. (2013). Efficient human pose estimation from single depth images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(12):2821–2840.

Sivic, J. y Zisserman, A. (2003). Video google: A text retrieval approach to object matching in videos. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 1470–1477. IEEE.

Stork, J. A., Spinello, L., et al. (2012). Audio-based human activity recognition using non-markovian ensemble voting. In *RO-MAN, 2012 IEEE*, pages 509–514. IEEE.

Veeriah, V., Zhuang, N., et al. (2015). Differential recurrent neural networks for action recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4041–4049.

Wallraven, C., Caputo, B., et al. (2003). Recognition with local features: the kernel recipe. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 257–264. IEEE.

Wang, A. et al. (2003). An industrial strength audio search algorithm. In *ISMIR*, pages 7–13. Washington, DC.

Wang, H., Ullah, M. M., et al. (2009). Evaluation of local spatio-temporal features for action recognition. In *BMVC 2009-British Machine Vision Conference*, pages 124–1. BMVA Press.

Wang, J., Yang, J., et al. (2010). Locality-constrained linear coding for image classification. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 3360–3367. IEEE.

Weinland, D., Ronfard, R., et al. (2011). A survey of vision-based methods for action representation, segmentation and recognition. *Computer Vision and Image Understanding*, 115(2):224–241.

Willems, G., Tuytelaars, T., et al. (2008). An efficient dense and scale-invariant spatio-temporal interest point detector. In *European conference on computer vision*, pages 650–663. Springer.

Xia, L., Chen, C.-C., et al. (2012). View invariant human action recognition using histograms of 3d joints. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*, pages 20–27. IEEE.

Yan, L., Casperson, D., et al. (2011). Survey: Dimension reduction by pattern decomposition. In *Modelling, Identification and Control (ICMIC), Proceedings of 2011 International Conference on*, pages 69–74. IEEE.

Yilma, A. y Shah, M. (2005). Recognizing human actions in videos acquired by uncalibrated moving cameras. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 1, pages 150–157. IEEE.

Yilmaz, A. y Shah, M. (2005). Actions sketch: A novel action representation. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 984–989. IEEE.

Zhang, T. y Kuo, C.-C. J. (2001). Audio content analysis for online audiovisual data segmentation and classification. *IEEE Transactions on speech and audio processing*, 9(4):441–457.

Zhu, Y., Ming, Z., et al. (2007). Automatic audio genre classification based on support vector machine. In *Third International Conference on Natural Computation (ICNC 2007)*, volume 1, pages 517–521. IEEE.